ASA-1150

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Appl. No.: 10/649,663 Confirmation No.: 6928

Applicant: N. SHIMOZONO et al.

Filed: August 28, 2003

Title: SWITCH PROVIDED WITH CAPABILITY OF SWITCHING A PATH

TC/AU: 2665

Examiner: C.L. Davis

Customer No.: 24956

SUBMISSION OF CERTIFIED PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Applicants submit herewith a certified priority document of the corresponding Japanese Patent Application:

No. 2003-203454, filed July 30, 2003, for the purpose of claiming foreign priority under 35 U.S.C. § 119.

Applicants respectfully request that the priority document be submitted and officially considered of record.

Respectfully submitted,

John R. Mattingly
Registration No. 30,293
Attorney for Applicant

MATTINGLY, STANGER, MALUR & BRUNDIDGE, P.C.
1800 Diagonal Road, Suite 370
Alexandria, Virginia 22314
(703) 684-1120
Date: January 24, 2006

10/647,663

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　７月３０日

出願番号
Application Number: 特願２００３－２０３４５４

[ST. 10/C]: [ＪＰ２００３－２０３４５４]

出願人
Applicant(s): 株式会社日立製作所

CERTIFIED COPY OF
PRIORITY DOCUMENT

２００３年　８月２６日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫

【書類名】　特許願
【整理番号】　K03009051A
【あて先】　特許庁長官殿
【国際特許分類】　G06F 13/00
【発明者】
【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
【氏名】　下薗　紀夫
【発明者】
【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
【氏名】　岩見　直子
【発明者】
【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
【氏名】　本田　聖志
【特許出願人】
【識別番号】　000005108
【氏名又は名称】　株式会社　日立製作所
【代理人】
【識別番号】　100075096
【弁理士】
【氏名又は名称】　作田　康夫
【手数料の表示】
【予納台帳番号】　013088
【納付金額】　21,000円
【提出物件の目録】
【物件名】　明細書　1
【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　パス切替えを提供するスイッチ

【特許請求の範囲】

【請求項１】

記憶装置及び計算機と接続されるスイッチであって、

前記記憶装置とは複数の通信線で接続されており、

前記記憶装置又は前記計算機と接続される複数のインターフェース部及び前記複数のインターフェース部を接続する内部スイッチを有し、

前記複数のインターフェース部のうちの第一のインターフェース部は、前記計算機からコマンドを受信して前記複数の通信線のうちの第一の通信線を介して前記記憶装置へ前記コマンドを転送し、前記第一の通信線の障害を検出したら、前記第一のインターフェース部は、前記計算機へ前記受信したコマンドのエラーを通知するフレームを送信し、前記障害を通知するフレームを前記計算機に送信した後は、前記第一のインターフェース部は、前記第一の通信線とは異なる第二の通信線を使用して前記計算機から受信するコマンドを前記記憶装置へ転送することを特徴とするスイッチ。

【請求項２】

前記第一のインターフェース部は、前記記憶装置へ送信したコマンドに対する応答を一定時間経過しても受信しない場合に、前記第一の通信線の障害と判断することを特徴とする請求項１記載のスイッチ。

【請求項３】

前記第一のインターフェース部は、前記記憶装置との物理的接続の切断を電気信号で判断して前記第一の通信線の障害を検出することを特徴とする請求項１記載のスイッチ。

【請求項４】

前記第一のインターフェース部は記憶部を有し、

前記第一のインターフェース部は前記コマンドを受信したら前記記憶部に前記コマンドを識別するための識別子を前記記憶部に記録し、前記コマンドで指定される処理の終了を示すフレームを前記記憶装置から受信したら、前記終了を示す

フレームに対応する前記コマンドの識別子を前記記憶部から消去し、

前記第一の通信線の障害を検出した場合、前記第一のインターフェース部は、その時点で前記記憶部に記録されている識別子に対応するコマンドのエラーを示すフレームを前記計算機に送信することを特徴とする請求項１記載のスイッチ。

【請求項５】

前記第一のインターフェース部は、前記記憶部に記録された識別子に対応する前記コマンドに基づいたデータの転送が開始された場合に、そのデータ転送の実施を示す情報を前記記憶部に記録し、

前記第一の通信線の障害を検出した場合、前記第一のインターフェース部は、前記記憶部に記録された識別子に対応するコマンドのうち、前記データの転送実施を示す情報が記録されている前記コマンドについて、エラーを示すフレームを作成し、前記計算機に送信することを特徴とする請求項４記載のスイッチ。

【請求項６】

前記第一のインターフェース部は、前記第一の通信線の障害を検出した場合、前記記憶部に記録された識別子に対応する前記コマンドのうち、前記データの転送の実施を示す情報が登録されていないコマンドについては、前記第二の通信線を介して前記コマンドを前記記憶装置へ送信することを特徴とする請求項５記載のスイッチ。

【請求項７】

前記第一のインターフェース部は、前記計算機へ仮想的な記憶装置を提供し、前記計算機から前記仮想的な記憶装置へのコマンドを受信した前記第一のインターフェース部は、前記仮想的な記憶装置へのコマンドを前記記憶装置へのコマンドへ変換することを特徴とする請求項１記載のスイッチ。

【請求項８】

記憶装置及び計算機と接続されるスイッチにおけるフレーム転送方法であって、

前記計算機からコマンドを受信して第一の通信線を介して前記記憶装置へ前記コマンドを転送し、

前記第一の通信線の障害を検出し、

前記計算機へ前記受信したコマンドのエラーを通知するフレームを送信し、

前記障害を通知するフレームを前記計算機に送信した後は、前記第一の通信線とは異なる第二の通信線を使用して前記計算機から受信するコマンドを前記記憶装置へ転送することを特徴とするフレーム転送方法。

【請求項９】

前記受信したコマンドを示す識別子を記録し、

前記第一の通信線の障害を検出したら、前記記録された識別子に対応するコマンドのエラーを通知するフレームを前記計算機に送信することを特徴とする請求項８記載のフレーム転送方法。

【請求項１０】

前記記録された識別子に対応するコマンドに基づくデータ転送の有無を記録し、

前記第一の通信線の障害を検出したら、前記記録された識別子に対するコマンドのうち、前記データ転送があるコマンドのエラーを通知するフレームを前記計算機に送信することを特徴とする請求項９記載のフレーム転送方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ネットワークで使用されるスイッチ又はネットワークに接続された装置に仮想化された記憶装置を提供するスイッチ（以下「仮想化スイッチ」）及びこれを含むコンピュータシステムに関する。

【０００２】

【従来の技術】

計算機（以下「コンピュータ」）と記憶装置システムとの間のデータ転送の信頼性を高めるための方法として、コンピュータと記憶装置システムとの間のアクセス経路（以下「アクセスパス」）を複数用意し、障害発生時にはアクセスパスを切替えることでコンピュータから記憶装置システムへのアクセスを継続するという技術が知られている。

【０００３】

例えば特許文献１には、複数の接続用ポートを持つ記憶装置ユニットを制御する記憶装置システムが、記憶装置ユニットへの１つのアクセスパスに障害が発生すると、他のアクセスパスを用いてコンピュータからのアクセス処理を継続し、コンピュータに対して障害の発生を隠蔽する技術が開示されている。

【０００４】

また特許文献２には、記憶装置システムが、自身が有する記憶装置へのアクセスパス障害を検出し、コンピュータに他のアクセスパスの使用を指示するエラー応答を返し、コンピュータが他のアクセスパスにアクセス要求（以下「コマンド」）を発行しなおすことで、アクセスパス切替えを実現する技術が開示されている。

【０００５】

【特許文献１】

特開平１１－１２００９２号公報

【特許文献２】

特開平１０－２１０１６号公報

【０００６】

【発明が解決しようとする課題】

特許文献１に開示された装置ではコンピュータに対してアクセスパスの障害を隠蔽することが出来るが、障害の隠蔽を実現するには、転送中のデータを装置内で一時的に保持するバッファ等の記憶装置が必須となる。これは、障害の隠蔽を図る装置が、障害発生時に行っていたデータ転送を他のアクセスパス（以下「代替パス」）を用いて再実行（以下「リトライ」）する必要があるためである。

【０００７】

ここで、特許文献１に開示された技術を適用する装置として、RAIDのような記憶装置システムを考える。この場合、記憶装置システムには性能向上を目的としてキャッシュメモリが搭載されることが多いので、このキャッシュメモリをリトライ用のコマンドやデータを保持するバッファとして使用すればよい。

【０００８】

しかし、最近コンピュータと記憶装置システムとを接続するストレージエリア

ネットワーク（以下「SAN」）と呼ばれる技術が出現しており、SANを構成しデータを中継する装置（以下「スイッチ」）にアクセスパス切替えの機能を持たせたいという要求がある。特に、スイッチが、複数の記憶装置システムの記憶領域を仮想化して一つ又は複数の仮想的な記憶装置システムをコンピュータに提供する場合、スイッチと実際の記憶装置システムとの間のアクセスパスの管理はスイッチで行わざるを得ない。

【０００９】

この場合、特許文献１に開示された技術をスイッチに適用しようとすると、余分なメモリをスイッチに持たせることなり、コスト増となる。

【００１０】

一方、特許文献２に開示された装置はコンピュータにアクセスパスの障害発生を通知し、コンピュータが代替パスにコマンドを再発行することでアクセスパス切替えを行う。従って、本装置では、コンピュータにアクセスパス切替えの機能を追加する必要がある。この場合でも、ネットワークに接続されたコンピュータ全てにアクセスパス切り替えのソフトウエア等をインストールする必要があるので、やはりコスト増となる。

【００１１】

本発明の目的は、コンピュータに機能を追加せずバッファなどのリソースを節約したアクセスパス切替えを可能にする装置および方法を提供することである。

【００１２】

【課題を解決するための手段】

本発明の一実施形態は以下に示す通りである。記憶装置とコンピュータを接続するスイッチが、パス切替えを行う。

【００１３】

具体的には、スイッチはコンピュータからコマンドを受信すると、コマンドを発行したコンピュータとコマンドの発行先である記憶装置などの対応情報を一時的に記憶する。その対応情報は、コマンドが完了する時に消去される。

【００１４】

又、スイッチはコマンドの発行先の記憶装置または記憶装置への経路で障害が

発生したことを判断し、対応情報に基づいてコンピュータへコマンドのエラー応答を返すとともに、以後の当該記憶装置へのコマンドの送信を代替パスで行うように設定を変更する。コンピュータはエラー応答を受信すると、同一の記憶装置にコマンドをリトライする。スイッチは、受信したコマンドを代替パスに中継する。これによりパス切替えが完了する。

【００１５】

尚、別の実施形態として、スイッチが仮想的な記憶装置を提供し、その仮想的な記憶装置へコンピュータがコマンドを送信したときにパス切り替えを行う構成が考えられる。

【００１６】

又、別の実施形態として、コマンドのエラーをコンピュータに報告する際、既にデータ転送の処理が開始されているコマンドについてのみエラーを報告する構成が考えられる。

【００１７】

【発明の実施の形態】

図１は、本発明を適用したコンピュータシステムの第一の実施形態の例を示す図である。

コンピュータシステムは、記憶装置104、コンピュータ（以下「ホスト」）105及びこれらを相互に接続するSAN101とを有する。

【００１８】

記憶装置104は、単体のハードディスクドライブ等を有する記憶装置や、複数の記憶装置を有する記憶装置システムである。又、ホスト105は一般的な計算機である。

【００１９】

SAN101は、記憶装置104とホスト105との間を接続し、ホスト105が記憶装置104に格納されたデータにアクセスするために使用するネットワークである。この様なSAN101は、例えばファイバチャネルやIPネットワークにSCSIプロトコルに従うフレームを流し、スイッチを用いてそのフレームをルーティングするように構成される。SAN101は、仮想化スイッチ102、スイッチ103を有し、仮想化スイッチ10

2には管理端末106が接続される。又、スイッチ間は通信線で相互に結合されている。

【００２０】

記憶装置104と仮想化スイッチ102との間は、接続しているケーブルの切断などの障害に対応する目的で、二系統のアクセスパスで接続されている。

【００２１】

仮想化スイッチ102、スイッチ103には、お互いを識別する為にSAN101内でユニークな識別子であるドメインアドレスが割り振られている。さらにSAN101に接続される各装置（記憶装置104及びホスト105）には、各々ユニークな識別子であるSANアドレスが割り振られる。SANアドレスは、SANアドレスが割り振られる装置と他のスイッチを介さずに直接接続される仮想化スイッチ102又はスイッチ103に割り振られているドメインアドレス及びそのスイッチに接続される装置群（以下「ドメイン」）内でユニークな識別子であるローカルアドレスとから構成される。尚、SANアドレスは、一つの装置に複数、例えばSAN101との物理的な接続ポートごとに割り振られても良い。尚本実施形態では、記憶装置104とホスト105はそれぞれ２つのSANアドレスを有するとする。

【００２２】

記憶装置104とホスト105は、コマンドやデータ等をフレームと呼ばれる単位で交換する。各フレームはヘッダ情報を格納する領域とペイロードから構成される。ヘッダ情報には送信元と送信先のSANアドレス（以下、「送信元アドレス」「送信先アドレス」）等が含まれる。又、ペイロードには、LUNと後述するコマンド、データ、レスポンスなどの情報が格納される。尚、LUNとは、記憶装置内部の論理的な記憶領域（以下「論理ユニット」）を表す番号である。

【００２３】

仮想化スイッチ102及びスイッチ103は、フレームの送信先アドレスに基づいてフレームの中継（「フレームルーティング」）を行う。

【００２４】

また、仮想化スイッチ102は、他のスイッチ103と同様にフレームルーティングを行うのに加え、ホスト101に仮想的な記憶装置（以下「仮想記憶装置107」）を

提供する。仮想記憶装置107は、実際は仮想化スイッチ102に接続される記憶装置104が有する幾つかの論理ユニットの集合体である。仮想記憶装置107が有する記憶領域に対するアクセスを記憶装置104が有する論理ユニットへのアクセスへ変換するために、仮想化スイッチ102は、仮想化記憶装置107の記憶領域と記憶装置104の論理ユニットとの対応関係に関する情報を有している。

【００２５】

ホスト105は、CPU1051、RAM1052、ホストバスアダプタ（以下「HBA」）1053を有する。CPU1051は、HBA1053を介してコマンドをペイロードに含むフレーム（以下「コマンドフレーム」）を記憶装置104や仮想記憶装置107に送信することで、記憶装置104等にコマンドを発行する。その際、ホスト105は、ヘッダ情報に先述した送信元アドレス、送信先アドレスに加えて、タグ（以下「ホストタグ」）を付加する。

【００２６】

ホストタグは、ホスト105が同一の記憶装置104へ複数のコマンドを発行した場合に、記憶装置104とホスト105との間で転送されるデータとホスト105が発行したコマンドとの対応関係を識別するために用いられる情報である。

【００２７】

記憶装置104は、２個のポート1041、コントローラ1042、メディア1043から構成される。ポート1041からコマンドフレームを受信した記憶装置104は、コントローラ1042でコマンドを解釈し、ポート1041を用いてデータをペイロードに含むフレーム（以下「データフレーム」）を送受信することでデータ転送を実行するとともにメディア1043のデータを読み書きする。尚、メディア1043は、ハードディスク、DVD等の不揮発性の記憶媒体であり、又、RAID等の冗長構成となっていても良い。

【００２８】

又、記憶装置104は、データ転送が完了すると、コマンドを発行した相手にコマンドが正常に終了したかどうかを示すレスポンスを返すことでデータ転送の完了をホスト104へ通知する。レスポンスは、レスポンス情報をペイロードに含むフレーム（以下「レスポンスフレーム」）を用いて通知される。データフレーム

やレスポンスフレームは、ヘッダ情報に送信元アドレス、送信先アドレスに加えて、データやレスポンスに対応するコマンドフレームに含まれていたホストタグを含む。

【００２９】

データフレームやレスポンスフレームを受信したホスト105は、そのフレームのホストタグをチェックすることでどのコマンドに対応するデータないしレスポンスであるのかを判断できる。

【００３０】

また、管理端末106はホスト105と同様の構成を持つ計算機である。管理者ないしユーザは、管理端末106を介して仮想化スイッチ102の設定を行うことが出来る。

【００３１】

図２は、仮想化スイッチ102の構成を示す図である。

仮想化スイッチ102は、複数のポートユニット111、管理ユニット112及びそれらを結合する内部スイッチ113を有する。

【００３２】

ポートユニット111は、SAN101に接続される各装置（以下「ノード」）であるスイッチ102、記憶装置104及びホスト105と仮想化スイッチ102とを接続するために使用されるインタフェイスである。ポートユニット111は、他のノードと接続されてフレームの送受信を行うSANインタフェイス1111、プログラムを実行するCPU1112、データを格納するRAM1113、プログラムを格納するROM1114、一定周期でCPU1112に信号を送るタイマ1115及び内部スイッチ113と接続される内部インタフェイス1116を有する。SANインタフェイス1111は、接続されたノードからフレームを受信すると、RAM1113内部のフレーム転送バッファ129に受信したフレームを書き込む。

【００３３】

管理ユニット112は、管理端末106と仮想化スイッチ102とを接続するために使用されるインタフェイスである。管理ユニット112は、管理端末106と接続され、管理用のコマンドやデータを送受信する管理インタフェイス1121、プログラムを

実行するCPU1122、データを格納するRAM1123、プログラムを格納するROM1124、後述するデータを格納するNVRAM1125及び内部スイッチ113と接続される内部インタフェイス1116を有する。

【００３４】

内部スイッチ113は、各ポートユニット111と管理ユニット112とを接続し、フレームや管理用メッセージの中継などを行う。管理用メッセージは、ポートユニット111と管理ユニット112の間で構成情報などを交換するためのメッセージである。

【００３５】

ポートユニット111のCPU1112と管理ユニット112のCPU1122とは、内部インタフェイス1116を介して、お互いのフレーム転送バッファ129にフレームの転送をしたり、管理用メッセージをお互いに送受信したりできる。

【００３６】

各ポートユニット111及び管理ユニット112は、お互いを識別するために仮想化スイッチ102内部でユニークなユニットIDという識別子を有する。内部インタフェイス1116は、CPU1112、1122によって指定されるユニットIDに対応するポートユニット111ないし管理ユニット112にフレームや管理用メッセージを転送する。

【００３７】

尚、上述した構成は仮想化スイッチ102の構成の一例であり、同等の機能を提供出来るならば、仮想化スイッチ102の構成は、本構成に限定されるものではない。例えばポートユニット111においてCPU1112、RAM1113、ROM1114、タイマ1115を、それらの機能を併せ持つ高機能プロセッサで置き換えても問題ない。

【００３８】

以下、本実施形態では、ホスト105が仮想記憶装置107にアクセスするものとして、仮想化スイッチ102が仮想化スイッチ102と記憶装置104との間のアクセスパスに発生する障害を検出し、障害の復旧を行う方法を説明する。

【００３９】

本発明の仮想化スイッチ102の動作の概要は以下のとおりである。

仮想化スイッチ102は、ホスト105からホストタグつきのコマンドフレームを受

け取り、そのホストタグを記録しかつコマンドフレームが指定する仮想記憶装置107に対応する記憶装置104を決定し、記憶装置104にタグ（仮想化スイッチ102が別途作成したタグ）つきのコマンドフレームを送信する。

【００４０】

記憶装置104はその応答として、タグつきのデータフレームやレスポンスフレームを返す。仮想化スイッチ102は受信したフレームを仮想記憶装置107からのリプライとしてホストタグを付加してホスト105に返す。この際、仮想化スイッチ102は、リプライされるフレームに関連するホストタグを記録から消去する。

【００４１】

さらにここで仮想化スイッチ102と記憶装置104間でアクセスパスの障害が発生した場合、仮想化スイッチ102は記録してあったホストタグに対応するレスポンスフレームをホスト105へ送信することでエラーを通知して、ホスト105にそのホストタグと関連するコマンドフレームを再送させる。それとともに、仮想化スイッチ102は、障害が発生したアクセスパスの切り替えを行い、以後ホスト105から仮想記憶装置107へのコマンドフレームを受信すると、他のアクセスパスを用いて記憶装置104へコマンドフレームを送信する。

【００４２】

以下、仮想化スイッチ102の処理手順の詳細を説明するが、その前に、仮想化スイッチ102での処理に必要なプログラム及びデータの構成について予め説明する。

【００４３】

図３は、ポートユニット111のROM1114に格納されるプログラム及びデータの構成を示す図である。ここに示されたプログラムはCPU1112で実行される。

ROM1114には、初期化プログラム121、ルーティングプログラム122、フレーム転送プログラム123、実仮想変換プログラム125及び障害処理プログラム126が格納されている。

【００４４】

初期化プログラム121は仮想化スイッチ102の起動時にCPU1112で実行され、CPU1112は、後述するルーティングテーブル128、フレーム転送バッファ129、コマン

ド管理テーブル130、アクセスパステーブル131及び仮想記憶装置構成テーブル132を初期化する。

【００４５】

ルーティングプログラム122は、SANインタフェイス1111や内部インタフェイス1116によってフレーム転送バッファ129にフレームが書き込まれるときにCPU1112が後述するルーティング処理を行い、フレームの転送先のユニットIDを決定する際に実行される。またルーティングプログラム122は、処理対象フレームが仮想記憶装置107のSANアドレスを送信先アドレスに含む場合は実仮想変換プログラム125を呼び出す処理を行う。

【００４６】

フレーム転送プログラム123は、フレーム転送バッファ129が保持しているフレームを、他のポートユニット111や管理ユニット112へ転送するためにCPU1112が実行するプログラムである。

【００４７】

実仮想変換プログラム125は仮想記憶装置107に対して発行されたコマンドを実行するための処理をCPU1112が行う際に実行されるプログラムで、コマンド開始処理1251、データフレーム処理1252及びコマンド終了処理1253のサブプログラムを有する。

【００４８】

障害処理プログラム126は、仮想化スイッチ102から記憶装置104に発行したコマンドの実行に障害が発生した場合、それを検出し障害を復旧する処理を行う際に、CPU1112で実行されるプログラムである。障害処理プログラム126は、タイムアウト検出処理1261及びリカバリ処理1262のサブプログラムを有する。タイマ1115からの信号に基づいてCPU1112はタイムアウト検出処理1261を定期的に実行し、仮想化スイッチ102から記憶装置104へ発行したコマンドのタイムアウトを検出する。タイムアウトとは、仮想化スイッチ102が記憶装置104にコマンドを発行してから、そのコマンドに対応するレスポンスを受信するまでの時間が一定以上になることを示す。タイムアウトを検出したCPU1112はリカバリ処理1262を実行し、障害を復旧する。

【００４９】

ルーティングテーブル128は、CPU1112がルーティング処理を行う際に必要となるテーブルである。テーブルの内容については後述する。

フレーム転送バッファ129は、RAM1113に確保された記憶領域で、CPU1112がフレームを処理するために、一時的にフレームを格納するためのバッファである。

【００５０】

コマンド管理テーブル130は、CPU1112が実仮想変換プログラム125及び障害処理プログラム126を実行する際に使用するテーブルである。

アクセスパステーブル131は、仮想記憶装置107を提供するために仮想化スイッチ102が使用する記憶装置104を、アクセスパスごとに管理するためのテーブルである。

仮想記憶装置構成テーブル132は、仮想記憶装置107とアクセスパスの対応を管理するテーブルである。

【００５１】

図４は、管理ユニット112のROM1124及びNVRAM1125に格納されたプログラム及びデータの構成を示す図である。

【００５２】

ROM1124には、初期化プログラム141、管理端末応答プログラム143及びフレーム転送プログラム123が格納されている。

【００５３】

初期化プログラム141は、仮想化スイッチ102の起動時にCPU1122で実行され、CPU1122は、ルーティングテーブル128とフレーム転送バッファ129を初期化し、仮想記憶装置構成テーブル132の内容に基づき、ルーティングテーブル128に仮想記憶装置107に対応するエントリを登録する。

【００５４】

管理端末応答プログラム143は、管理インタフェイス1121を介した管理端末106からの要求に応じて、CPU1122がアクセスパステーブル131、仮想記憶装置構成テーブル132の変更を行い、それに伴いルーティングテーブル128を変更する際に実行される。

【００５５】

尚、NVRAM1123には、ポートユニット111と同じ内容のアクセスパステーブル131、仮想記憶装置構成テーブル132、ルーティングテーブル128及びフレーム転送バッファ129が格納されている。

【００５６】

図５は、アクセスパステーブル131の構成例を示す図である。

アクセスパステーブル131は、複数のアクセスパスエントリ1311を有する。

【００５７】

１つのアクセスパスエントリ1311には、仮想記憶装置107の提供に用いられる記憶装置104の記憶領域への１つのアクセスパスに関する情報が登録される。具体的には、個々のアクセスパスエントリ1311は、管理端末106によって設定されるアクセスパスIDが格納されるフィールド1312、記憶装置104のSANアドレスとLUNから構成される実SANアドレスの情報が登録されるフィールド1313、記憶装置104のLUN（以下「実LUN」）の情報が登録されるフィールド1314及び当該エントリ1131で指定される記憶装置104へのアクセスパスが利用可能であるかを示すステータスの情報が登録されるフィールド1315の各フィールドを有する。

尚、フィールド1315に登録される情報の値が１であれば、当該エントリで指定されるアクセスパスは使用可能であり、０であれば使用不能であることを示す。

【００５８】

図６は、仮想記憶装置構成テーブル132の構成例を示す図である。

仮想記憶装置構成テーブル132は、複数の仮想記憶装置エントリ1321からなる。一つの仮想記憶装置エントリ1321は、仮想記憶装置107のSAN101への接続、具体的には、仮想記憶装置107に付与されるSANアドレスに対応している。

【００５９】

個々の仮想記憶装置エントリ1321は、仮想記憶装置107に付与されるSANアドレス（以下「仮想SANアドレス」）の情報が登録されるフィールド1322、仮想記憶装置が有する論理ユニットの番号（以下「仮想LUN」）を示す情報が登録されるフィールド1323、正アクセスパスを示すID（以下「正アクセスパスID」）が登録されるフィールド1324、副アクセスパスを示すID（以下「副アクセスパスID」）

が登録されるフィールド1325及び転送先ユニットを示すID（以下「転送先ユニットID」）を登録するフィールド1326を有する。

【００６０】

仮想SANアドレスは、仮想化スイッチ102のドメインアドレスと管理端末106によって設定されるローカルアドレスから構成される。仮想LUNは、管理端末106によって設定される。

正アクセスパスは、通常時に仮想化スイッチ102が使用する記憶装置104へのアクセスパスを表し、副アクセスパスは、障害発生時に仮想化スイッチ102が使用する記憶装置104へのアクセスパスを示す。これら双方のアクセスパスのIDは、管理端末106によって設定される。

【００６１】

転送先ユニットIDは、当該エントリ1321に対応する仮想記憶装置107に対するコマンドを実行するポートユニット111を示すユニットIDを表し、管理端末106によって設定される。

【００６２】

図８は、ルーティングテーブル128の構成例を示す図である。

ルーティングテーブル128は、複数のルーティングエントリ1281を有する。

各ルーティングエントリ1281は、フレームに含まれる送信先アドレスと、そのフレームを転送すべきポートユニット111のユニットIDとの対応関係を表す。各々のルーティングエントリ1281は、送信先アドレスである宛先SANアドレスを登録するフィールド1282、フィールド1282に登録された送信先アドレスを有するフレームの転送先となるポートユニット111ないし管理ユニット112を表す転送先ユニットIDの情報が登録されるフィールド1283及び当該エントリに対応するフレームの送信先アドレスが仮想記憶装置107に対応するかを示す仮想フラグが登録されるフィールド1283を有する。

【００６３】

フィールド1283に登録される仮想フラグが１であれば、当該エントリは仮想記憶装置107に対応するSANアドレスに対応し、０であればポートユニット111のSANインタフェイス1116に接続されたノードに対応する。管理ユニット112は、仮想

記憶装置構成テーブル132に登録されたエントリに対応する宛先SANアドレス及び転送先ユニットIDの組を、仮想フラグを１にしてルーティングテーブル128に追加する。

【００６４】

図１０は、コマンド管理テーブル130の構成例を示す図である。

コマンド管理テーブル130は、複数のコマンド管理エントリ1301を有する。各コマンド管理エントリ1301は、ホスト105から仮想記憶装置107へ発行された各コマンドに対応している。各々のコマンド管理エントリ1301は、ホストSANアドレスを登録するフィールド1302、仮想SANアドレスを登録するフィールド1303、仮想LUNを登録するフィールド1304、コマンドに含まれるホストタグの情報を登録するフィールド1305、使用アクセスパスIDを登録するフィールド1306、生成タグを登録するフィールド1307及びタイムアウトカウンタとして使用されるフィールド1308の各フィールドを有する。

【００６５】

ホストSANアドレスとは、コマンドを発行したホスト105のSANアドレスを表す。フィールド1303には、コマンドフレームの送信先となる仮想記憶装置107の仮想SANアドレスの情報が登録される。フィールド1304には、コマンドフレームで指定された仮想記憶装置107のLUNの情報が登録される。

【００６６】

フィールド1305には、ホスト105が発行したコマンドフレームのヘッダに含まれるホストタグの値が登録される。使用アクセスパスIDとは、後述するコマンド開始処理で決定されるアクセスパスIDである。生成タグとは、仮想化スイッチ102が記憶装置104に対して発行するコマンドフレームに使用されるタグであり、後述するコマンド登録処理で生成される。

【００６７】

タイムアウトカウンタは、後述するタイムアウト検出処理が、当該コマンドのタイムアウトを検出するために用いる値である。またタイムアウトカウンタの値が－１である場合、そのコマンド管理エントリ1301は無効であることを示す。

【００６８】

以下、仮想化スイッチ102におけるフレームの処理手順について説明する。仮想化スイッチ102は、受信したフレームを所定のノードにルーティングする際に、そのフレームの宛先が仮想記憶装置107であるかどうかを判断して、実仮想変換等の処理を行う。尚、以下の説明では、仮想化スイッチ102にて実仮想変換を行うポートユニット111は予め決まっており、ルーティングテーブル128に登録される仮想SANアドレスに対応する転送先ユニットIDは、その実仮想変換を行うポートユニット111を指定するユニットIDであるとする。

【００６９】

図９は、実仮想変換処理を行うポートユニット111におけるルーティング処理フローを示す図である。

当該処理は、ポートユニット111のSANインタフェイス1111がフレームを受信しフレーム転送バッファ129に書き込んだとき、ないし他のポートユニット111からフレームが転送されたときに実行される。

【００７０】

まずポートユニット111は、フレームの送信先アドレスと宛先SANアドレス1282が一致するルーティングエントリ1281をルーティングテーブル128から選択する（ステップ152）。

【００７１】

次にポートユニット111は、ステップ152で選び出したルーティングエントリ1281のフィールド1283に登録された転送先ユニットIDが、本処理を実行しているポートユニット111のユニットIDと一致しているか判定する（ステップ153）。

【００７２】

一致する場合、ポートユニット111は、ステップ152で選び出されたルーティングエントリ1281のフィールド1283に登録された仮想フラグの値が１であるか判定する（ステップ154）。仮想フラグの値が１の場合、ポートユニット111は、実仮想変換プログラム125を実行して、フレームの転送先を仮想記憶装置107から記憶装置104又はホスト105へ変換し、宛先を変換したフレームを記憶装置104又はホスト105へ転送する。具体的には、ポートユニット111は、宛先となる記憶装置104又はホスト105と接続されている他のポートユニット111へフレームを転送する

。実仮想変換処理の詳細は後述する（ステップ155）。

【００７３】

又、ステップ153で、フィールド1283に登録された転送先ユニットIDが本処理を実行しているポートユニット111のユニットIDと一致しない場合、ポートユニット111は、フレームの転送先をステップ152で選び出されたルーティングエントリ1281のフィールド1283に登録された転送先ユニットIDに設定する（ステップ158）。

【００７４】

又、ステップ154で仮想フラグが０であった場合、ポートユニット111は、SANインタフェイス1111を用いて処理対象フレームを宛先SANアドレスで指定された装置へ送信し、処理を終了する（ステップ159）。

【００７５】

又、ステップ158の処理後、ポートユニット111は、フレーム転送プログラム123を実行して、処理対象フレームをステップ158で設定された転送先ユニットIDに基づいて、他のユニットに処理対象フレームを転送して、処理を終了する（ステップ160）。

【００７６】

尚、ポートユニット111が実仮想変換処理を行わないポートユニット111である場合には、上述した処理においてステップ154、155の処理は行われず、ポートユニット111は、ステップ153でフレームが自ポート宛であると判断したら、ステップ159の処理を実行する。

【００７７】

次に、ポートユニット111で実行される実仮想変換処理の詳細について説明する。実仮想変換処理は、ホスト105から仮想記憶装置107へのコマンドフレームを記憶装置104へのコマンドフレームへ変換するコマンド開始処理、ホスト105又は記憶装置104から受信したデータフレームの宛先を変換するデータフレーム処理、記憶装置104から受信したレスポンスフレームの宛先を変換するコマンド終了処理の３種類の処理に分けられる。上述したルーティング処理のステップ155で実仮想変換処理を開始したポートユニット111は、受信したフレームの内容に応

じて、上述した３つの処理を行う。以下、３つの処理手順の詳細について説明する。

【００７８】

図１１は、ポートユニット111で行われるコマンド開始処理1251の処理フローを示す図である。

まずポートユニット111は、受信したコマンドフレームの送信先アドレスと仮想SANアドレスが一致する仮想記憶装置エントリ1321を仮想記憶装置構成テーブル132から選択する（ステップ161）。

【００７９】

次に、ポートユニット111は、ステップ161で選び出した仮想記憶装置エントリ1321のフィールド1324及び1325に登録された正アクセスパスと副アクセスパスのどちらを使用するか選択する。具体的には、ポートユニット111は、登録された正アクセスパスに対応するアクセスパスエントリ1311をアクセスパステーブル131から選び出し、そのエントリ1311のフィールド1315に登録されたステータスをチェックする。ステータスが１なら、その正アクセスパスを選択する。ステータスが１でなければ、副アクセスパスに対応するアクセスパスエントリ1311をアクセスパステーブル131から選択する（ステップ162）。

【００８０】

その後、ポートユニット111は、ステップ162で選び出したアクセスパスエントリ1311に対応する記憶装置104に対して発行するコマンドに付加すべきタグ（以下「生成タグ」）を生成する。この生成タグには、コマンド管理テーブル130の有効ないずれのコマンド管理エントリ1301のフィールド1307に登録された生成タグのいずれとも一致しない値が用いられる（ステップ163）。

【００８１】

その後、ポートユニット111は、コマンド管理テーブル130に処理対象となるコマンドを登録する。具体的には、ポートユニット111は、無効なコマンド管理エントリ1301をコマンド管理テーブル130から１つ選択し、そのエントリ1301の各フィールド1302、1303、1304、1305、1306及び1307に、処理対象コマンドフレームに含まれる送信元アドレス、送信先アドレス、LUN、ホストタグ、ステップ162

で選択されたアクセスパスエントリ1311のアクセスパスID、ステップ163で生成された生成タグ値を設定し、更に、フィールド1308のタイムアウトカウンタを０に設定する（ステップ164）。

【００８２】

次にポートユニット111は、処理対象となるコマンドフレームのヘッダ情報とペイロードに含まれるLUNを書き換える。具体的には、ポートユニット111は、コマンドフレームに含まれる送信元アドレス、送信先アドレス、LUN及びホストタグの値を、処理対象のコマンドフレームの送信先アドレス（つまり仮想記憶装置107のSANアドレス）、ステップ162で選択されたアクセスパスエントリ13111の実SANアドレス、実LUN及びステップ163で生成された生成タグ値に書き換える（ステップ165）。

【００８３】

その後、ポートユニット111は、ルーティングプログラム122を実行して、ステップ165で変換されたコマンドフレームを所定の記憶装置104へ転送し、処理を終了する（ステップ166）。

【００８４】

図１２は、ポートユニット111におけるデータフレーム処理1252の処理フローを示す図である。

まずポートユニット111は、受信したデータフレームの送信元アドレス、送信先アドレス及びタグが、フィールド1302、1303及び1305に登録された値と一致するコマンド管理エントリ1301をコマンド管理テーブル130から検索する（ステップ171）。

【００８５】

一致するコマンド管理エントリ1301を発見した場合、ポートユニット111は、選び出したコマンド管理エントリ1301のフィールド1306に登録された使用アクセスパスIDに対応するアクセスパスエントリ1311をアクセスパステーブル131から選び出す。そして、ポートユニット111は、受信したデータフレームのヘッダ情報を書き換える。具体的には、ポートユニット111は、受信したデータフレームの送信元アドレス、送信先アドレス及びタグを、ステップ171で選び出したコマ

ンド管理エントリ1301のフィールド1303に登録された仮想SANアドレス、本ステップで選び出したアクセスパスエントリ1311のフィールド1313に登録された実SANアドレス及びステップ171で選び出したコマンド管理エントリ1301のフィールド1307に登録された生成タグの値に書き換える（ステップ172）。

【００８６】

ステップ171で一致するコマンド管理エントリ1301を発見できなかった場合、ポートユニット111は、受信したデータフレームの送信先アドレス及びタグが、フィールド1303及び1307に登録された値と一致するコマンド管理エントリ1301をコマンド管理テーブル130から選択する（ステップ173）。

【００８７】

その後、ポートユニット111は、受信したデータフレームのヘッダ情報を書き換える。具体的には、データフレームの送信元アドレス、送信先アドレス及びタグを、ステップ171で選び出したコマンド管理エントリ1301のフィールド1303に登録された仮想SANアドレス、フィールド1302に登録されたホストSANアドレス及びフィールド1305に登録されたホストタグに書き換える（ステップ174）。

【００８８】

ステップ172又はステップ174の処理を終了したポートユニット111は、ルーティングプログラム122を実行して、ステップ172ないしステップ174で書き換えたデータフレームを所定のノード（ホスト105又は記憶装置104）に転送して処理を終了する（ステップ175）。

【００８９】

図１３は、ポートユニット111における、コマンド終了処理1253の処理フローを示す図である。

まずポートユニット111は、受信したレスポンスフレームの送信先アドレス及び生成タグがフィールド1303及びフィールド1307に登録された値と一致するコマンド管理エントリ1301をコマンド管理テーブル130から選択する（ステップ181）

その後、ポートユニット111は、レスポンスフレームのヘッダ情報及びLUNを書き換える。具体的には、ポートユニット111は、レスポンスフレームの送信元アドレス、送信先アドレス及び生成タグを、ステップ181で選択したコマンド管理

エントリ1301のフィールド1303に登録された仮想SANアドレス、フィールド1302に登録されたホストSANアドレス及びフィールド1305に登録されたホストタグに書き換える（ステップ182）。

【００９０】

その後、ポートユニット111は、ステップ181で選択したコマンド管理エントリ1301を無効化するために、フィールド1308のタイムアウトカウンタの値を－１に設定する（ステップ183）。

【００９１】

最後に、ポートユニット111は、ルーティングプログラム122を実行して、ステップ182で書き換えたレスポンスフレームをホスト105へ転送し、処理を終了する（ステップ184）。

以上の処理により、仮想化スイッチ102は、実仮想変換処理を行って、仮想記憶装置107が関連するフレームの処理を行うことが出来る。

【００９２】

次に、仮想化スイッチ102が仮想化スイッチ102と記憶装置104の間のアクセスパスを検出して代替パスの使用を行う間の処理手順を説明する。仮想化スイッチ102は、記憶装置104へ送信したコマンドに対する応答が一定時間経っても帰ってこないことを検出し、それを契機に代替パスを使用するための処理を行う。前者の処理をタイムアウト検出処理、後者の処理をリカバリ処理と称し、以下説明する。尚、上記二つの処理を実行するのは、上述した実仮想変換処理を行うポートユニット111であるとする。

【００９３】

図１４は、ポートユニット111で実行されるタイムアウト検出処理1281の処理フローを示す図である。CPU1112はタイマ1115からの信号に基づいて本処理を定期的に実行する。

まずポートユニット111は、コマンド管理テーブル130から先頭のコマンド管理エントリ1301を選び出す。尚、２回目以降にステップ191を実行する場合、ポートユニット111は、前回のコマンド管理エントリ1301の次のコマンド管理エントリ1301を選択する（ステップ191）。

【００９４】

次に、ポートユニット111は、ステップ191で選び出したコマンド管理エントリ1301が有効であるか判定する。具体的には、フィールド1308のタイムアウトカウンタの値が-1か否かでエントリ1301が有効かどうか判定する（ステップ192）。エントリ1301が有効な場合、ポートユニット111は、ステップ191で選び出したコマンド管理エントリ1301のフィールド1308のタイムアウトカウンタ値を１増やす（ステップ193）。

その後、ポートユニット111は、ステップ191で選び出したコマンド管理エントリ1301のフィールド1308のタイムアウトカウンタ値が、タイムアウトと判定される特定の定数値に等しいか判定する（ステップ194）。

【００９５】

ステップ194でタイムアウトと判定された場合、ポートユニット111は、リカバリ処理を実行する。リカバリ処理の詳細は後述する（ステップ195）。

ステップ192でエントリ1301が無効なエントリであると判定された場合、ステップ194でエントリ1301がタイムアウトしていないと判定された場合又はステップ195のリカバリ処理が終了した場合、ポートユニット111は、ステップ191で選択したコマンド管理エントリ1301がコマンド管理テーブル130の最後のエントリであるか判定する。最後のエントリならば処理を終了し、さもなくばステップ191へ戻って次のエントリ1301について処理を繰り返す（ステップ196）。

【００９６】

図１５は、ポートユニット111におけるリカバリ処理の処理フローを示す図である。

まず、ポートユニット111は、コマンド管理テーブル130から先頭のコマンド管理エントリ1301を選択する。尚、２回目以降にステップ191を実行する場合は、ポートユニット111は、前回のコマンド管理エントリ1301の次のコマンド管理エントリ1301を選択する（ステップ201）。

【００９７】

次にポートユニット111は、ステップ201で選択したコマンド管理エントリ1301が、図１３で説明したタイムアウト検出処理のステップ194でタイムアウトと判

断されたエントリ1301のフィールド1306に登録された使用アクセスパスIDと同一の使用アクセスパスIDを持つか判定する（ステップ202）。

選択されたエントリ1301が、タイムアウトと判断されたエントリ1301と同じ使用アクセスパスIDを有する場合、ポートユニット111は、ステップ201で選択したコマンド管理エントリ1301に対応するコマンドに対するエラー応答のためのレスポンスフレームを生成する。具体的には、ステップ201で選択されたコマンド管理エントリ1301のフィールド1303に登録された仮想SANアドレス、フィールド1302に登録されたホストSANアドレス及びフィールド1305に登録されたホストタグを、各々レスポンスフレームの送信元アドレス、送信先アドレス及びタグとする。レスポンスの内容はホスト105に対してコマンドのリトライを要請する内容にする（ステップ203）。

【００９８】

その後、ポートユニット111は、ステップ201で選択したコマンド管理エントリ1301を無効化するために、フィールド1308のタイムアウトカウンタの値を－１に設定する（ステップ204）。

エントリ1301を無効化した後、ポートユニット111は、ルーティングプログラム122を実行して、ステップ203で生成したレスポンスフレームをホスト105へ転送する（ステップ205）。

【００９９】

ステップ205の処理の終了後又はステップ202で選択されたエントリ1301がタイムアウトと判断されたエントリ1301と同じ使用アクセスパスIDを有しない場合、ポートユニット111は、ステップ201で選択されたコマンド管理エントリ1301がコマンド管理テーブル130の最後のエントリであるか判定する。最後のエントリでない場合には、ポートユニット111は、ステップ201以降の処理を繰り返す（ステップ206）。

【０１００】

選択されたエントリ1301がコマンド管理テーブル130の最後のエントリの場合、ポートユニット111は、図１３のステップ194で指定された使用アクセスパスIDとアクセスパスIDが一致するアクセスパスエントリ1311をアクセスパステーブル

131から選び出し、そのフィールド1315のステータスの値を０に設定する。そして処理を終了する（ステップ207）。

【０１０１】

本実施形態においては、仮想化スイッチ102と記憶装置104の間の物理的な接続が切断されるなどの理由で、正常時に使用する記憶装置104のSANアドレスへ送信したコマンドに対するデータ転送やレスポンスが記憶装置104から仮想化スイッチ102に送信されない時、仮想化スイッチ102が、そのコマンドのタイムアウトを検出しホスト105に対してコマンドのリトライを要請する内容のエラーを含むレスポンスを送信することが出来る。

【０１０２】

上述の処理と並行して、仮想化スイッチ102は、正アクセスパスIDに対応するアクセスパスエントリ1311を無効化するため、以後の仮想記憶装置107へのコマンドは副アクセスパスIDを用いて処理されるようになる。そのため、ホスト105が仮想記憶装置107に対してコマンドをリトライすると、仮想化スイッチ102は副アクセスパスIDを用いて、記憶装置104のもうひとつのSANアドレスへコマンドを送信する。これによりホスト105は仮想記憶装置107へ継続してアクセスを続行することが可能になる。

【０１０３】

尚、本実施形態では、仮想化スイッチ102がタイムアウトを検出することで仮想化スイッチ102と記憶装置104間のパス障害等を検出したが、この方法以外の方法を用いてもよい。例えば、仮想化スイッチ102と記憶装置104とが直接物理的に接続されているときには、仮想化スイッチ102は、電気信号の変化や光信号の消灯などで記憶装置104等の障害を検出することも可能である。この場合、障害を検出したポートユニット111は、管理メッセージを管理ユニット112へ送信し、管理ユニット112が各ポートユニット111へ管理メッセージを送信することで、各ポートユニット111は障害の発生を検出し、リカバリ処理を実行してもよい。

【０１０４】

また、本実施形態では仮想化スイッチ102が仮想記憶装置107を提供した。しかし、仮想記憶装置107と記憶装置104の１つのパスを一致させても問題はない。つ

まり仮想化スイッチ102は仮想記憶装置107を必ずしも提供しなくても良く、その場合でも上記のパス切替え処理を行うことは可能である。

【０１０５】

以下、本発明を適用したコンピュータシステムの第二の実施形態について説明する。第二の実施形態のコンピュータシステムの構成は第一の実施形態の構成と同様であるが、仮想化スイッチ102が以下に説明する本実施形態の仮想化スイッチ302に置き換わる点が異なる。

【０１０６】

第１の実施形態の仮想化スイッチ102は、タイムアウトを検出した時点で当該仮想記憶装置107に関する全てのコマンドに対してエラー応答をホスト105に通知していた。本実施形態の仮想化スイッチ302は、コマンドのタイムアウトを検出した時点で、データ転送が開始されてしまっているコマンドについてのみエラー応答をホスト105に返し、その他のコマンドについては仮想化スイッチ102が記憶装置104へ他の経路を使ってコマンドフレームの再送を行う。仮想化スイッチ302を利用すると、第一の実施形態に比べて、ホスト105がリトライする必要のあるコマンドの数を減らすことが可能になる。以下、第二の実施形態について、第一の実施形態と異なる部分のみ説明する。

【０１０７】

仮想化スイッチ302の構成は、仮想化スイッチ102の構成と同様である。しかし、コマンド開始処理1251、データフレーム処理1252及びリカバリ処理1262の手順並びにコマンド管理テーブル130の構成が若干異なる。

【０１０８】

図１５は、本実施形態におけるコマンド管理テーブル330の構成例を示す図である。

コマンド管理テーブル330は、複数のコマンド管理エントリ3301から構成される。コマンド管理エントリ3301は、第一の実施形態におけるコマンド管理エントリ1301に加えて、転送起動フラグを登録するフィールド3302及びコマンド内容を登録するフィールド3303を有する。

【０１０９】

転送起動フラグは、仮想化スイッチ102が当該コマンドに関するデータフレームを受信したことがあるかどうかを示すフラグである。転送起動フラグの値が０ならば、仮想化スイッチ102が過去にデータフレームを受信したことが無いことを、１ならば既に受信したことがあることを示す。

【０１１０】

コマンド内容は、ホスト105から仮想記憶装置107に対して送信されたコマンドフレームのペイロードに含まれるコマンドの内容である。

【０１１１】

本実施形態のコマンド開始処理1251は、上述したステップ183の処理において、ポートユニット111が、コマンド管理エントリ3301のフィールド3302のデータ転送フラグを０に設定し、コマンドの内容をフィールド3303に保存する処理が付け加わる。

【０１１２】

本実施形態のデータフレーム処理1252は、上述したステップ172、173の処理において、ポートユニット111が使用するコマンド管理エントリ3301のフィールド3302のデータ転送フラグを１に設定する処理が加わる。

【０１１３】

図１６は、本実施形態におけるポートユニット111のリカバリ処理1262の処理フローを示す図である。

まず、ポートユニット111は、図１４のステップ201と同様の処理を行う（ステップ401）。

【０１１４】

次に、ポートユニット111は、図１５のステップ202と同様の判定を行う（ステップ402）。

選択されたエントリ3301が、タイムアウトと判断されたエントリ3301と同じ使用アクセスパスIDを有する場合、ポートユニット111は、ステップ401で選び出したコマンド管理エントリ3301のフィールド3302に登録された転送起動フラグが１か判定する（ステップ403）。

【０１１５】

転送起動フラグが０の場合、ポートユニット111は、ステップ401で選択したコマンド管理エントリ3301を用いて、交替パスへ再発行するコマンドを生成する。具体的には、ポートユニット111は、まずコマンド管理エントリ3301のフィールド1303及び1304と一致するフィールド1322及び1323を持つ仮想記憶装置エントリ1321を仮想記憶装置構成テーブル132から選択する。そして、その選択されたエントリ1321のフィールド1325に登録された副アクセスパスIDと一致するアクセスパスIDを持つアクセスパスエントリ1311をアクセスパステーブル131から選び出す。

【０１１６】

そして、ポートユニット111は、ステップ401で選び出したコマンド管理エントリ3301のフィールド1302の仮想SANアドレス、ステップ404で選び出したアクセスパスエントリ1311のフィールド1313の実SANアドレス、フィールド1314の実LUN、ステップ401で選び出したコマンド管理エントリ3301のフィールド1307の生成タグ及びフィールド3303のコマンド内容を、生成するコマンドフレームの、送信元アドレス、送信先アドレス、LUN、生成タグ及びコマンドにする（ステップ404）。

【０１１７】

その後、ポートユニット111は、ステップ401で選び出したコマンド管理エントリ3301のフィールド1306の値を、ステップ404で選び出したアクセスパスエントリ1311のフィールドID1312に登録された値へ変更し、フィールド1308のタイムアウトカウンタを０に設定する（ステップ405）。

【０１１８】

一方、ステップ403で転送起動フラグが１だった場合、ポートユニット111は、図１４のステップ203及び204と同様の処理を行う（ステップ406、407）。

ステップ405の処理後又はステップ407の処理後、ポートユニット111は、ルーティングプログラム122を実行し、ステップ404で生成されたコマンドフレームないしステップ406で生成されたレスポンスフレームを記憶装置104又はホスト105へ転送する（ステップ408）。

【０１１９】

ステップ408の処理の終了後又はステップ402で選択されたエントリ3301が、タイムアウトと判断されたエントリ3301と同じ使用アクセスパスIDを有しないと判断された場合、ポートユニット111は、図１４のステップ206及び207と同様の処理を行い、処理を終了する（ステップ409、410）。

【０１２０】

本実施形態における仮想化スイッチ302と仮想化スイッチ102との違いは、データ転送が始まっていないコマンドに対するリカバリ処理の手順である。

具体的には、仮想化スイッチ102は、データ転送が始まっているかどうかに関わらず、タイムアウトと判定されたコマンドについて、エラー応答をホスト105に送信した。

【０１２１】

一方、本実施形態の仮想化スイッチ302は、各コマンドについてコマンド管理エントリ3301で、データ転送が始まっているかどうかを判断する転送起動フラグと、そのコマンドの内容を記録している。そのため、データ転送が始まっていないコマンドに対しては、仮想化スイッチ302は、ホスト105にエラーを返すことなく記憶装置104のもうひとつのSANアドレスへコマンドを再送する。これにより、ホスト105がリトライしなくてはならないコマンドの数を減らすことが可能になる。

【０１２２】

次に、本発明を適用したコンピュータシステムの第３の実施形態について説明する。

【０１２３】

本実施形態では、第１、第２の実施形態と同等の機能を提供する仮想化スイッチの異なる構成について説明し、仮想化スイッチ全体としてのリソースを削減可能であることを示す。

図１８は、本実施形態における仮想化スイッチ502の構成例を示す図である。仮想化スイッチ502と仮想化スイッチ102の違いは、ポートユニット511が、ポートユニット111からタイマ1115が省かれた構成である点と、新たにポートユニット111からSANインタフェイス1111を省いた構成を有する仮想化ユニット513が追

加される点である。

【０１２４】

ポートユニット511のROM1114には、ポートユニット111のROM1114に格納されているプログラムやデータのうち、実仮想変換プログラム125、障害処理プログラム126、コマンド管理テーブル130、アクセスパステーブル131及び仮想記憶装置構成テーブル132以外のプログラム等が格納されている。

また、仮想化ユニット513に格納されるプログラム及びデータの構成はポートユニット111の構成と同様である。

【０１２５】

本実施形態における仮想化スイッチ502では、仮想化ユニット513が仮想記憶装置107を提供する。このような構成、即ち通常のフレームルーティングの処理と仮想記憶装置107に関する処理を切り離す構成とすることで、ポートユニット511はポートユニット111よりもCPU1112の負荷やRAM1113の容量を節約でき、タイマ1115を取り除くことが出来る。その代わり、仮想化スイッチ502には仮想化ユニット513を追加する必要があるが、ポートユニット513を多く持つ仮想化スイッチ502の場合は、本構成の方が仮想化スイッチ102よりも全体のリソースを節約することが可能である。

【０１２６】

また、同様にして第２の実施形態の仮想化スイッチ302のリカバリ処理を本実施例の仮想化スイッチ502の構成で行うことも可能である。

尚、上述した各プログラムの実行により行われる処理は、専用のハードウエアで実現されても良い。

【０１２７】

【発明の効果】

本発明により、記憶装置とコンピュータを接続するネットワークにおいて、データ転送用のバッファやキャッシュを用いることなく、またコンピュータに機能を追加することなく、パス切替えを行うスイッチを提供することが可能となる。

【図面の簡単な説明】

【図１】

本発明を適用する第１の実施形態のコンピュータシステムの構成図である。

【図２】

第１の実施形態の仮想化スイッチの構成図である。

【図３】

第１の実施形態のポートユニットのプログラム及びデータの構成図である。

【図４】

第１の実施形態の管理ユニットのプログラム及びデータの構成図である。

【図５】

第１の実施形態のアクセスパステーブルの構成図である。

【図６】

第１の実施形態の仮想記憶装置構成テーブルの構成図である。

【図７】

第１の実施形態のルーティングテーブルの構成図である。

【図８】

第１の実施形態のルーティング処理のフローチャートである。

【図９】

第１の実施形態のコマンド管理テーブルの構成図である。

【図１０】

第１の実施形態のコマンド開始処理のフローチャートである。

【図１１】

第１の実施形態のデータフレーム処理のフローチャートである。

【図１２】

第１の実施形態のコマンド終了処理のフローチャートである。

【図１３】

第１の実施形態のタイムアウト検出処理のフローチャートである。

【図１４】

第１の実施形態のリカバリ処理のフローチャートである。

【図１５】

第２の実施形態のコマンド管理テーブルの構成図である。

【図１６】

第２の実施形態のリカバリ処理のフローチャートである。

【図１７】

第３の実施形態の仮想化スイッチの構成図である。

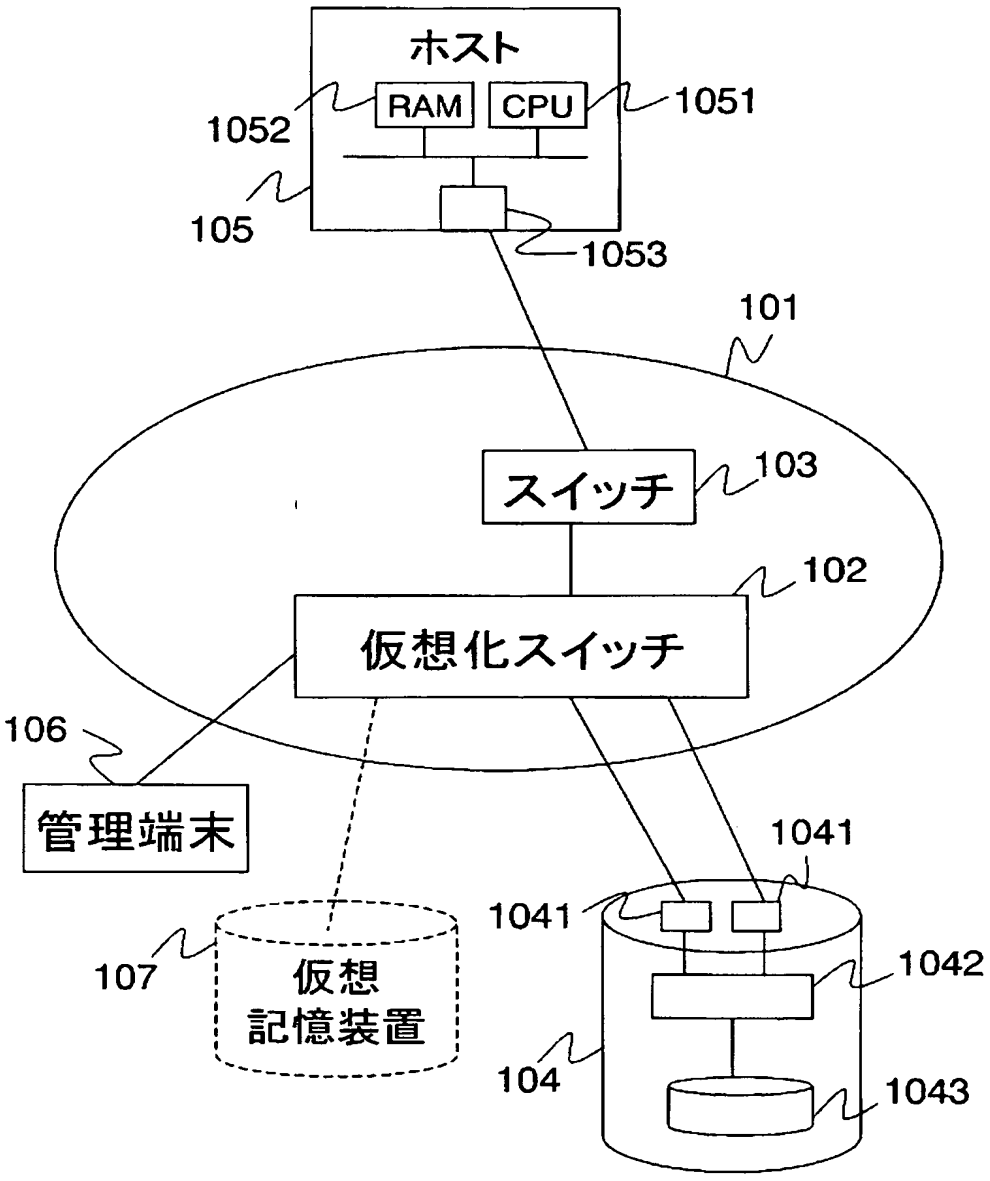
【符号の説明】

101…SAN、102…仮想化スイッチ、103…スイッチ、104…記憶装置、105…ホスト、106…管理端末、107…仮想記憶装置、111…ポートユニット、112…管理ノード。

【書類名】　図面

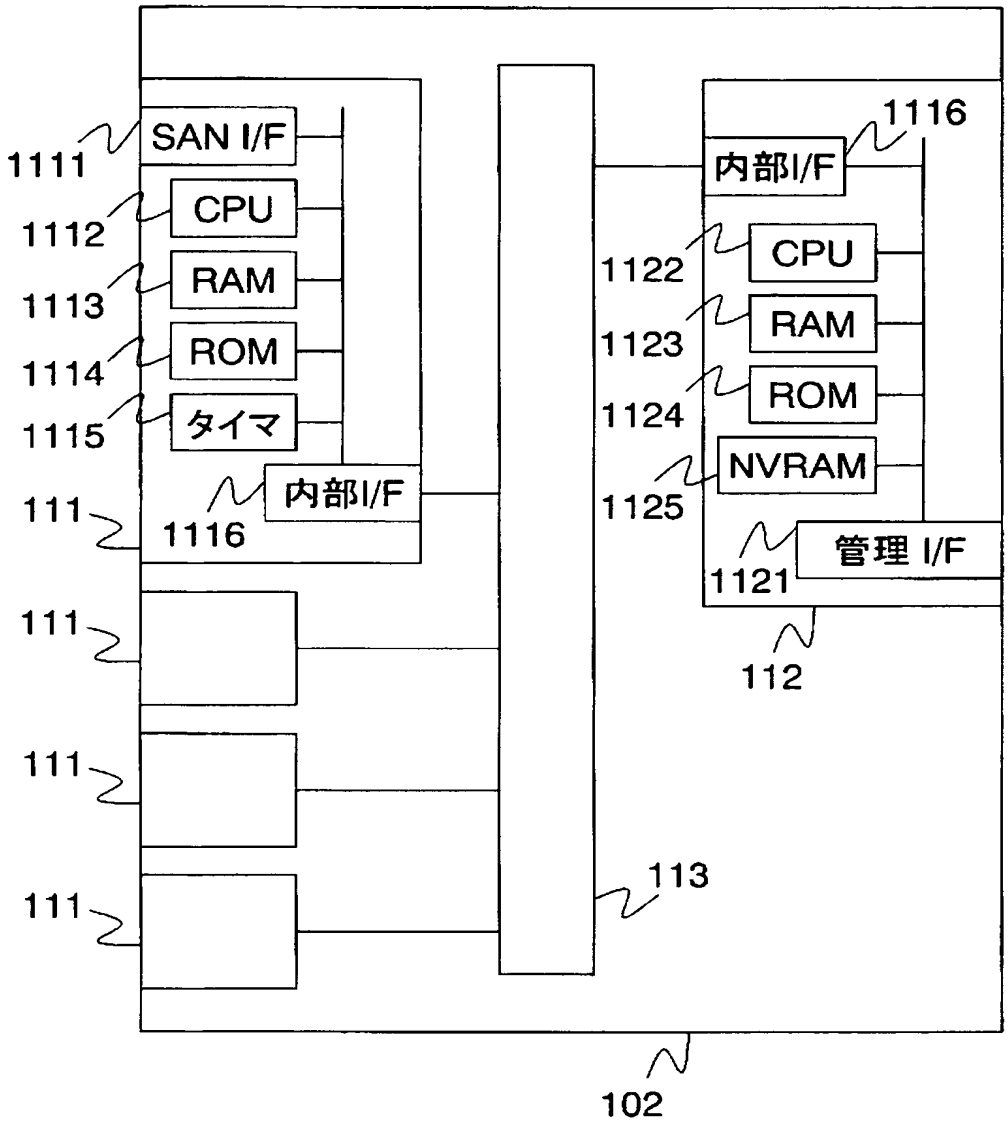
【図１】

図　1


ホスト
RAM
CPU
1052
1051
105
1053
101
スイッチ
103
102
仮想化スイッチ
106
管理端末
1041
1041
107
仮想
記憶装置
1042
104
1043

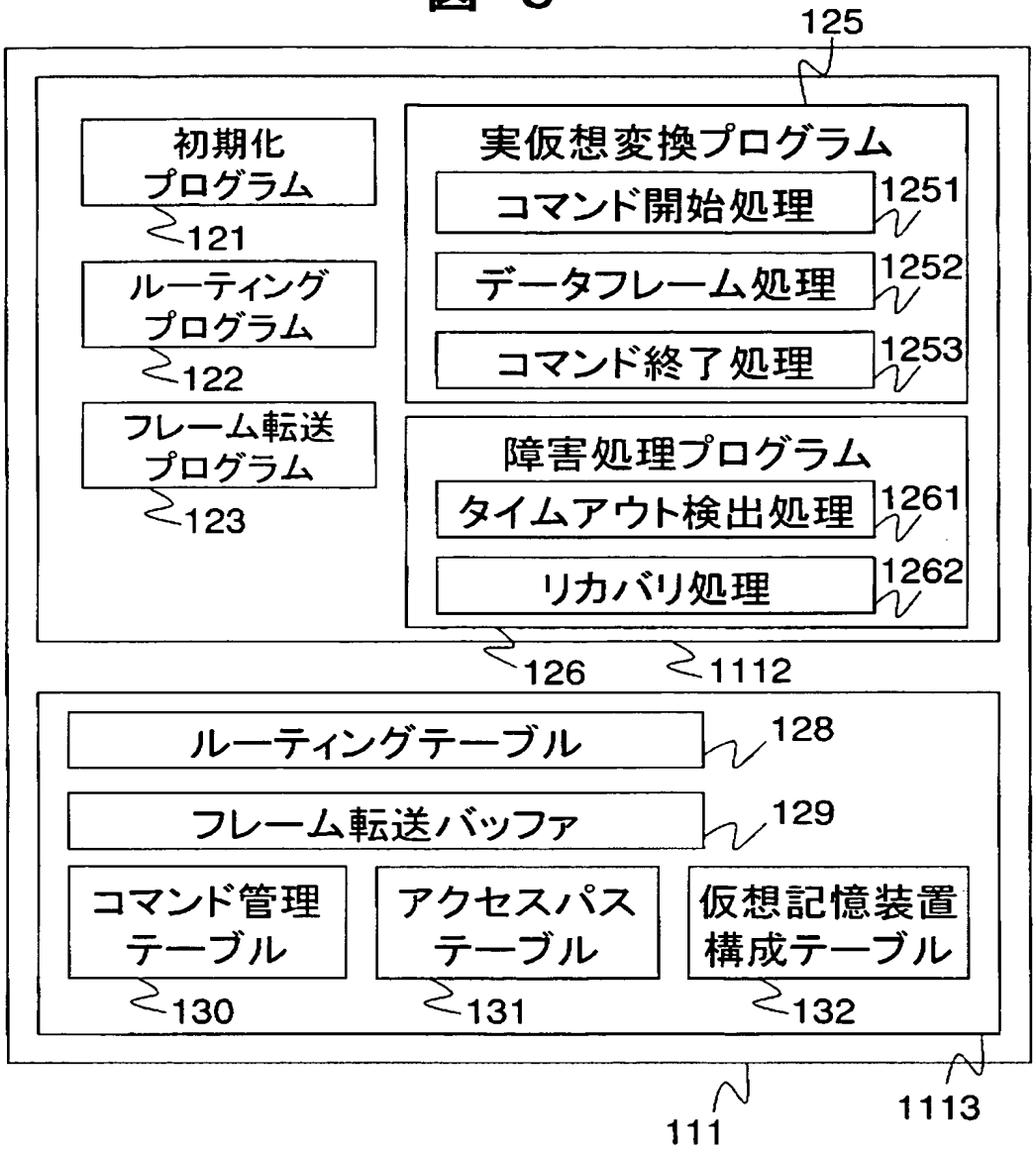
【図２】

図　2


SAN I/F
CPU
RAM
ROM
タイマ
内部I/F
内部I/F
CPU
RAM
ROM
NVRAM
管理 I/F
1111
1112
1113
1114
1115
1116
111
1116
1122
1123
1124
1125
1121
112
111
111
111
113
102

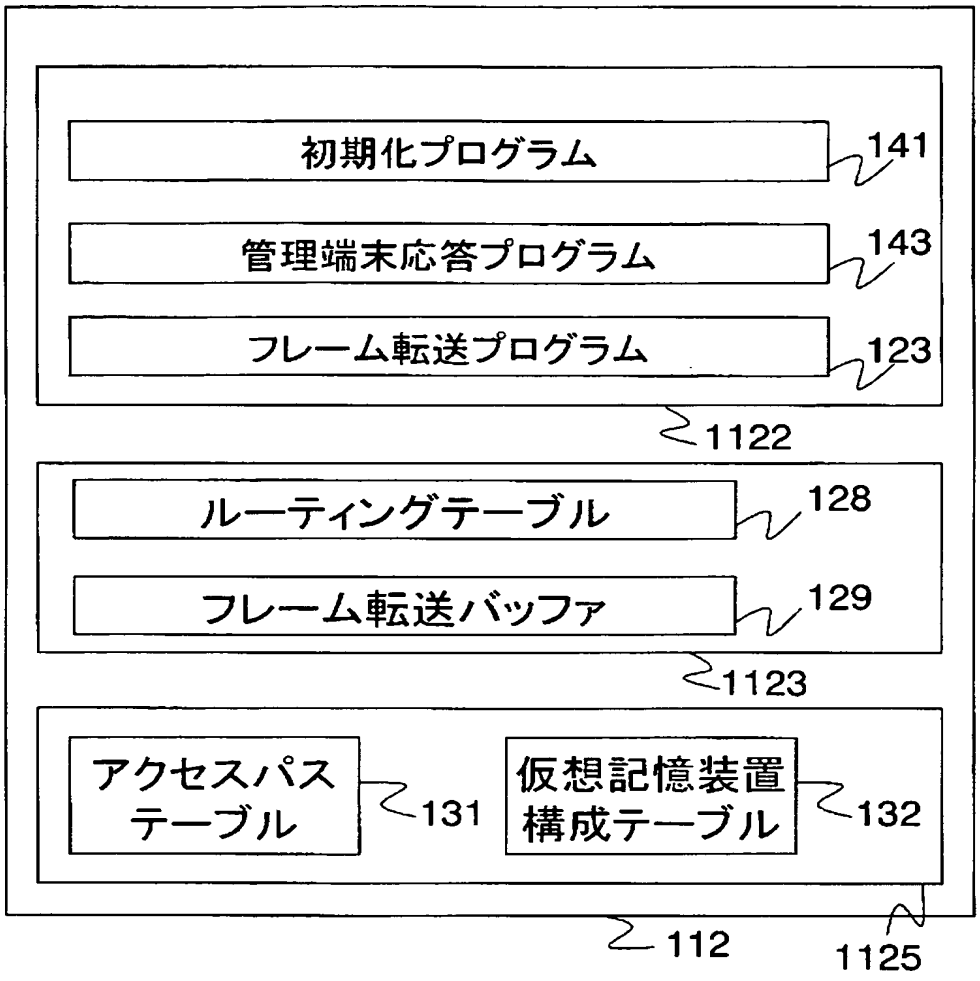
【図３】

図　3


125
初期化
プログラム
121
ルーティング
プログラム
122
フレーム転送
プログラム
123
実仮想変換プログラム
コマンド開始処理
1251
データフレーム処理
1252
コマンド終了処理
1253
障害処理プログラム
タイムアウト検出処理
1261
リカバリ処理
1262
126
1112
ルーティングテーブル
128
フレーム転送バッファ
129
コマンド管理
テーブル
130
アクセスパス
テーブル
131
仮想記憶装置
構成テーブル
132
111
1113

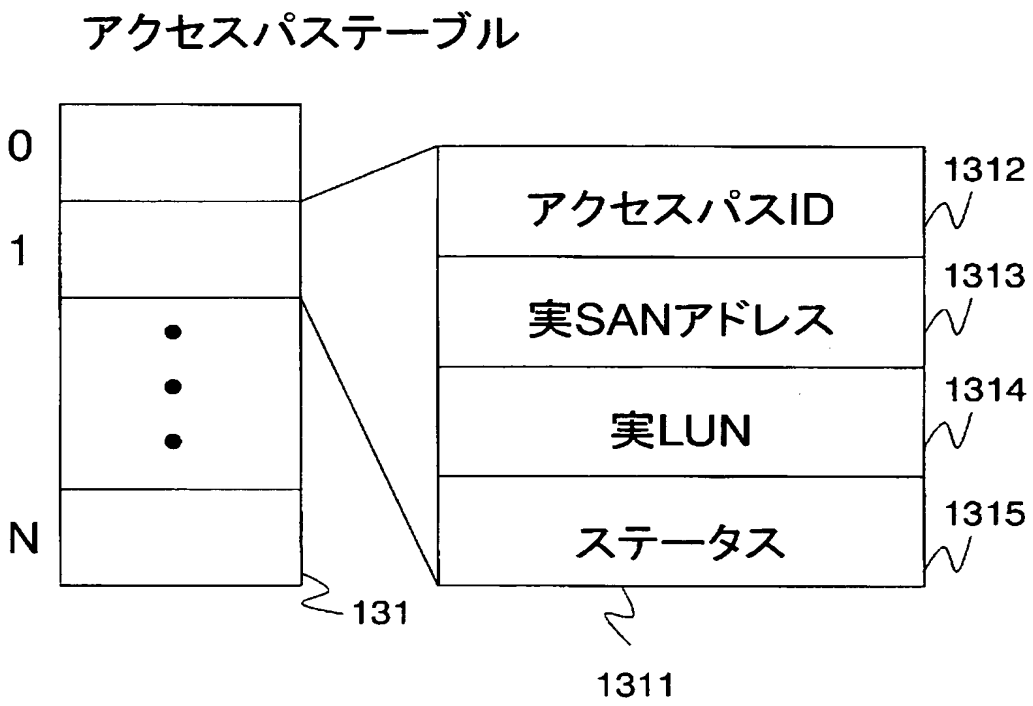
【図４】

図 4


初期化プログラム
141
管理端末応答プログラム
143
フレーム転送プログラム
123
1122
ルーティングテーブル
128
フレーム転送バッファ
129
1123
アクセスパス
テーブル
131
仮想記憶装置
構成テーブル
132
112
1125

【図５】

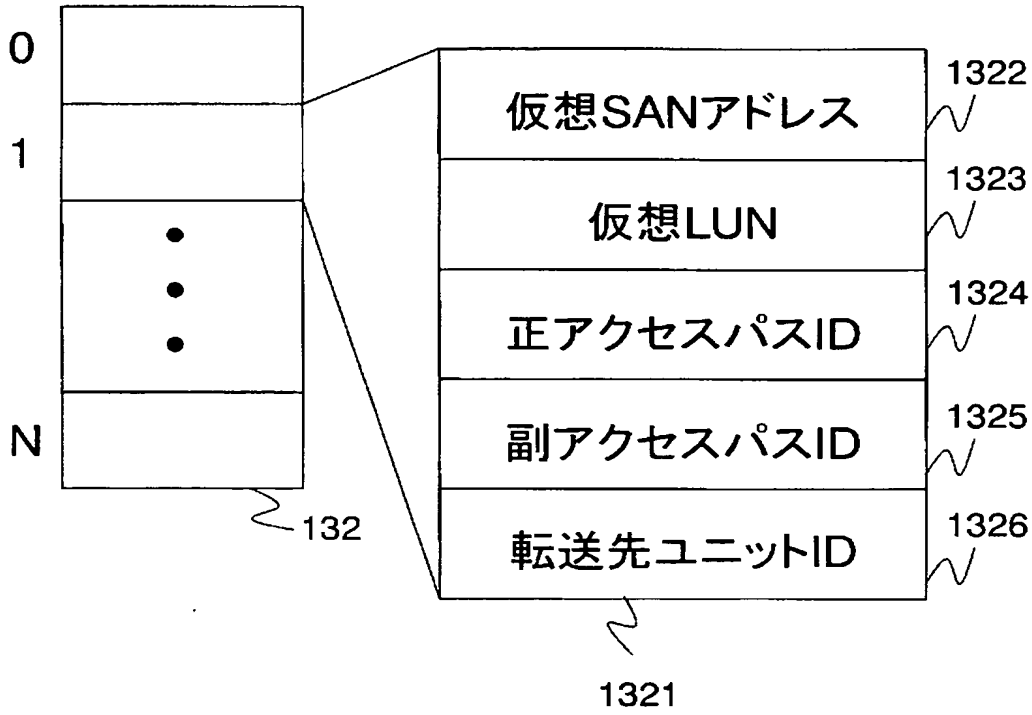
図 5


アクセスパステーブル
0
1
N
131
アクセスパスID
実SANアドレス
実LUN
ステータス
1312
1313
1314
1315
1311

【図６】

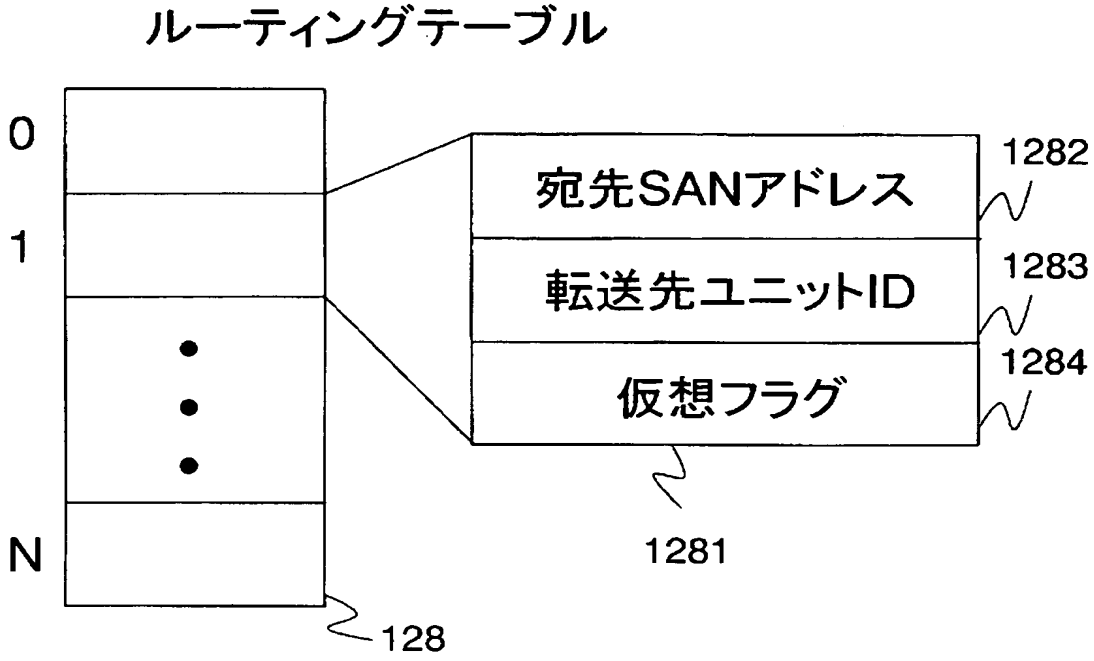
図　6

仮想記憶装置構成テーブル



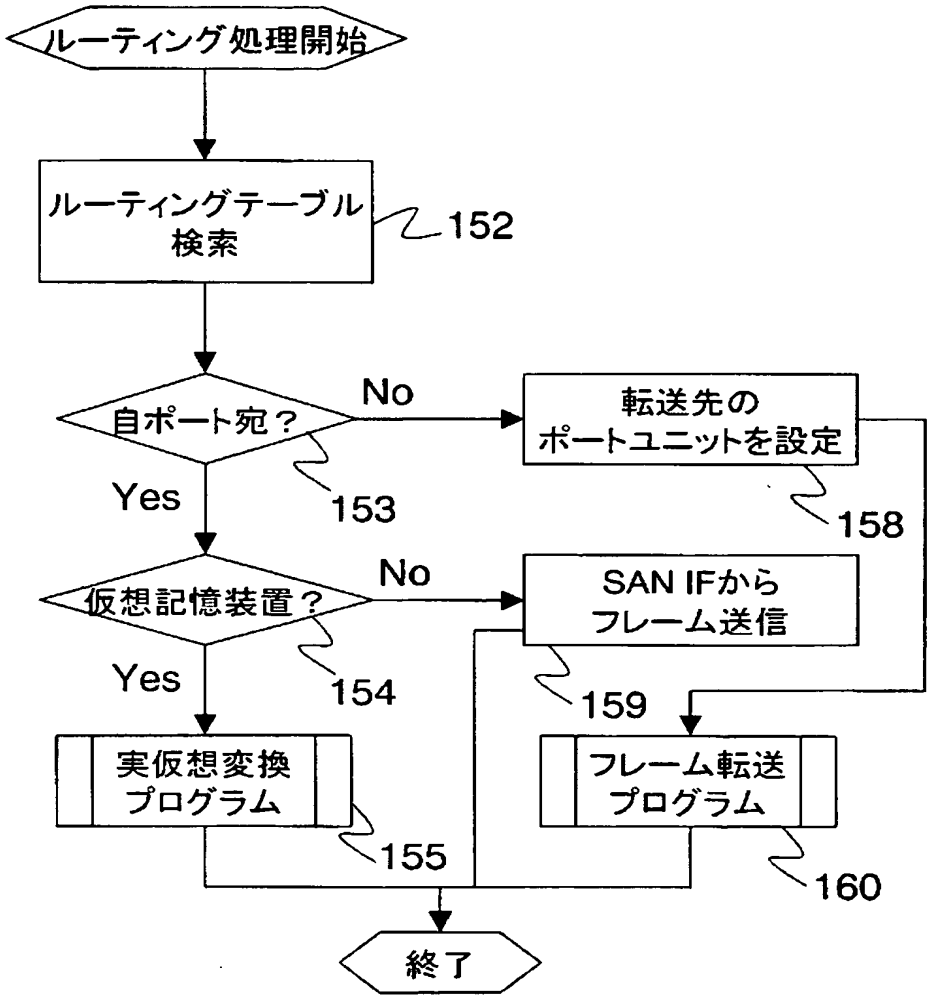
【図７】

図 7


ルーティングテーブル
0
1
N
128
宛先SANアドレス
転送先ユニットID
仮想フラグ
1281
1282
1283
1284

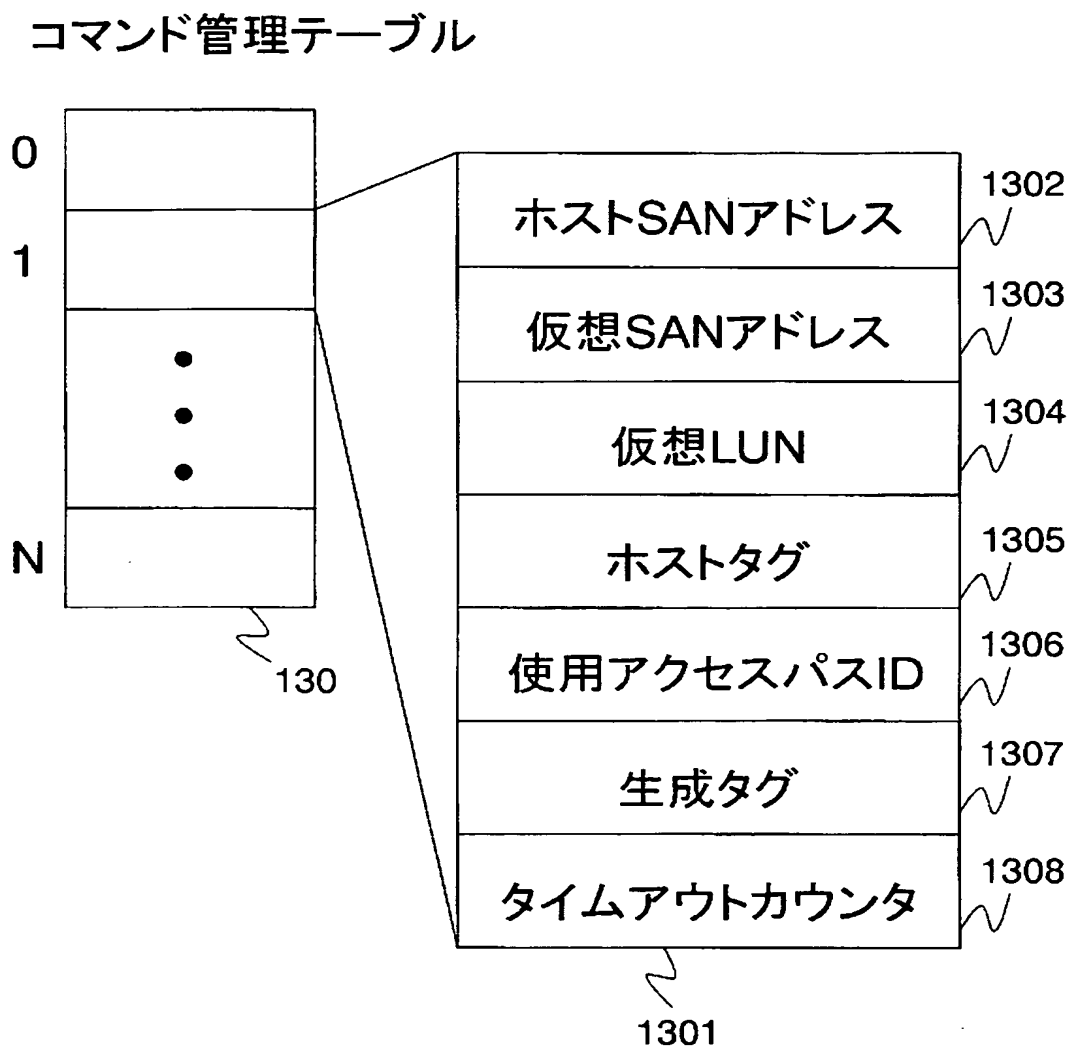
【図８】

図 8


ルーティング処理開始
ルーティングテーブル検索
152
自ポート宛？
No
153
転送先のポートユニットを設定
158
Yes
仮想記憶装置？
No
154
SAN IFからフレーム送信
159
Yes
実仮想変換プログラム
155
フレーム転送プログラム
160
終了

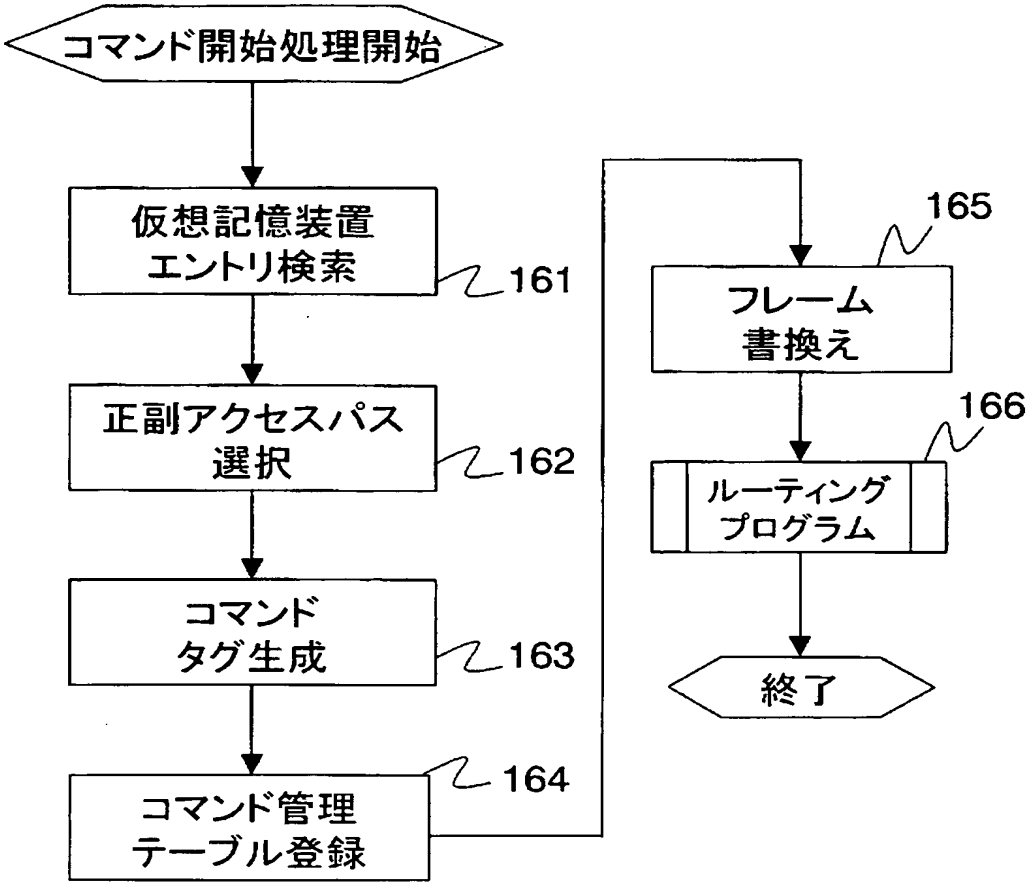
【図９】

図 ９


コマンド管理テーブル
0
1
N
130
ホストSANアドレス
仮想SANアドレス
仮想LUN
ホストタグ
使用アクセスパスID
生成タグ
タイムアウトカウンタ
1302
1303
1304
1305
1306
1307
1308
1301

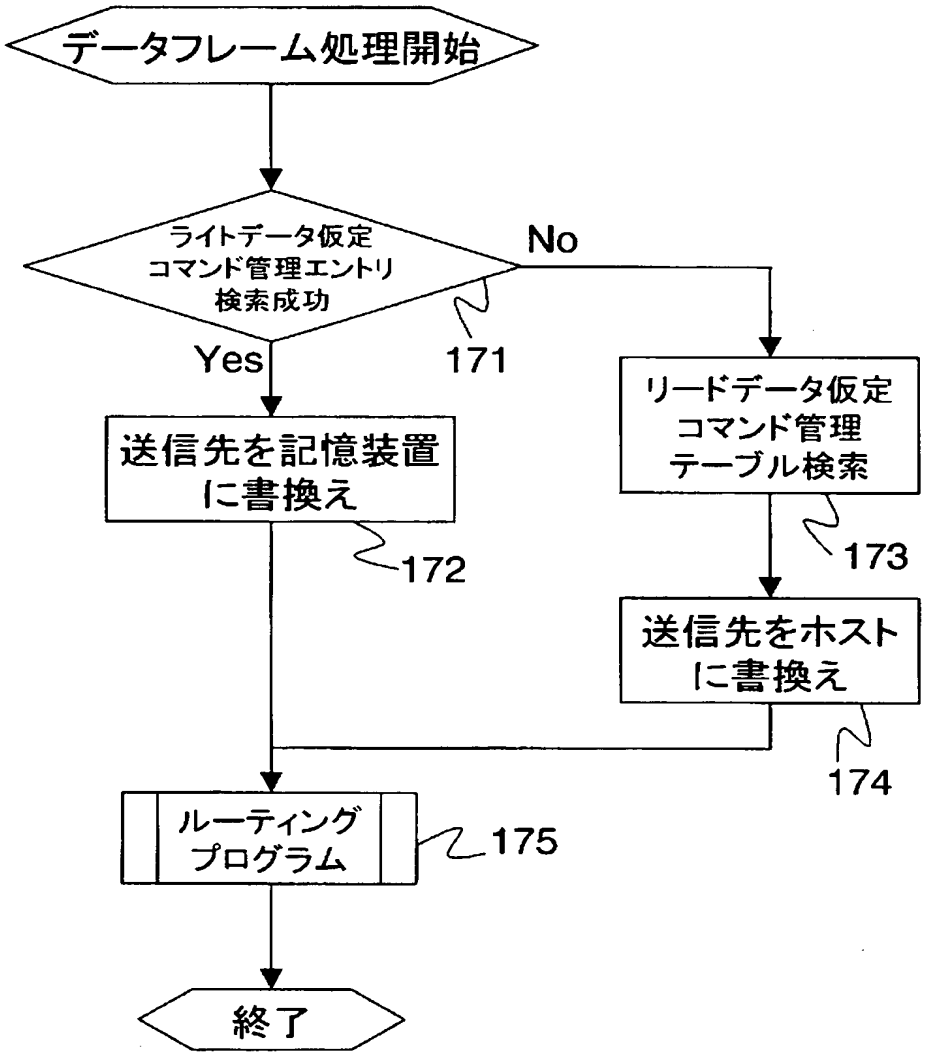
【図１０】

図 10


コマンド開始処理開始
仮想記憶装置
エントリ検索
161
正副アクセスパス
選択
162
コマンド
タグ生成
163
コマンド管理
テーブル登録
164
フレーム
書換え
165
ルーティング
プログラム
166
終了

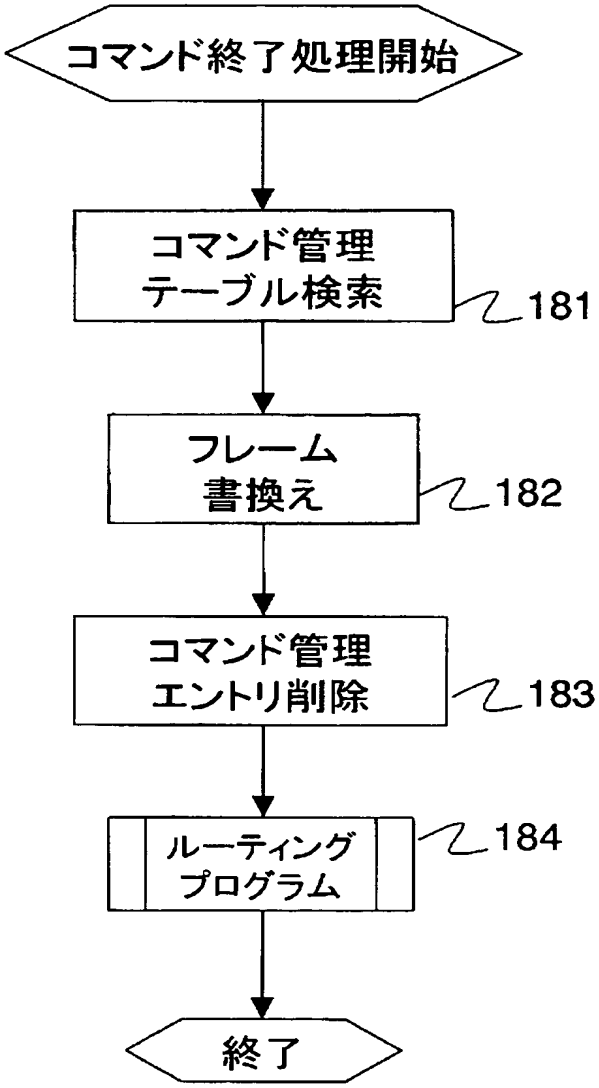
【図１１】

図 11


データフレーム処理開始
ライトデータ仮定
コマンド管理エントリ
検索成功
No
Yes
171
送信先を記憶装置
に書換え
172
リードデータ仮定
コマンド管理
テーブル検索
173
送信先をホスト
に書換え
174
ルーティング
プログラム
175
終了

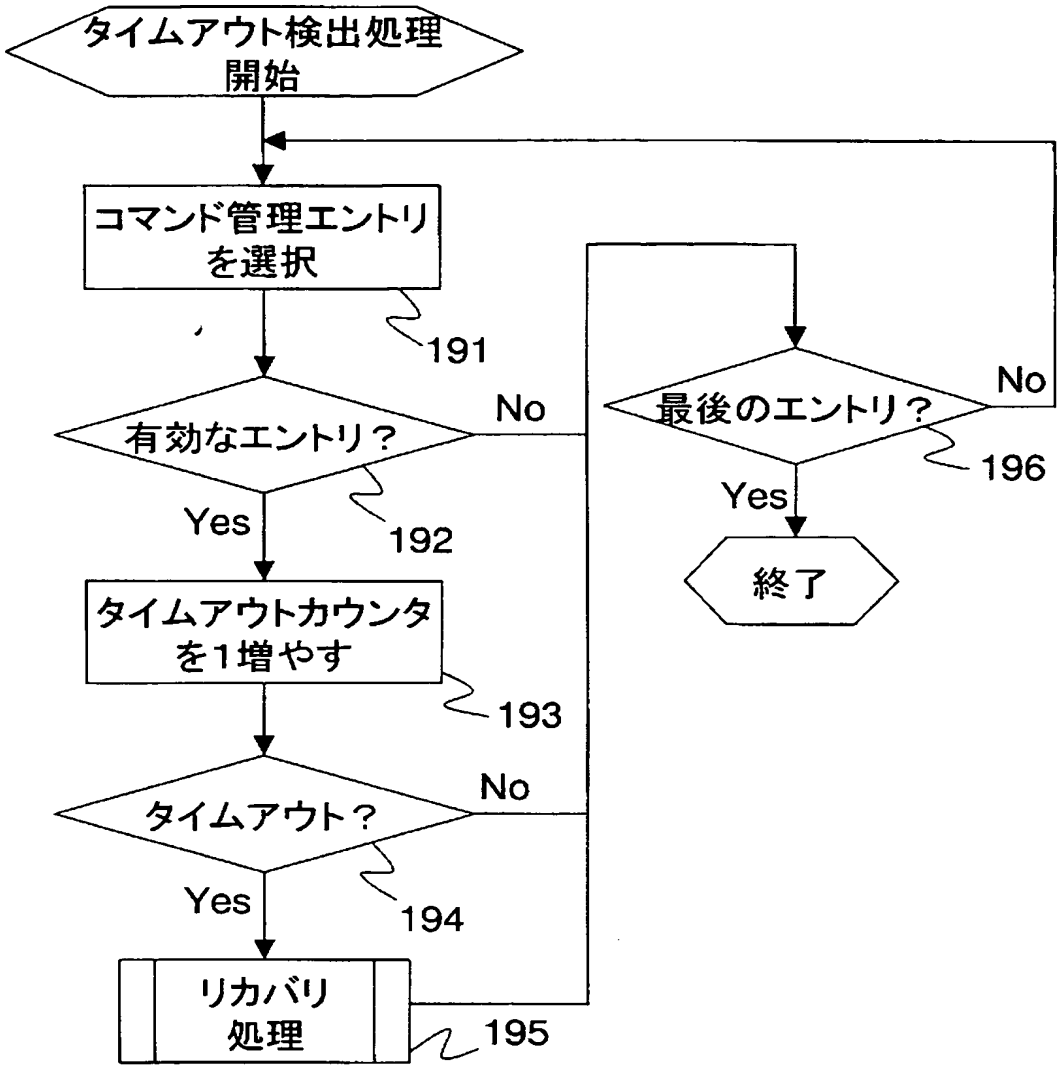
【図１２】

図 12


コマンド終了処理開始
コマンド管理
テーブル検索
181
フレーム
書換え
182
コマンド管理
エントリ削除
183
ルーティング
プログラム
184
終了

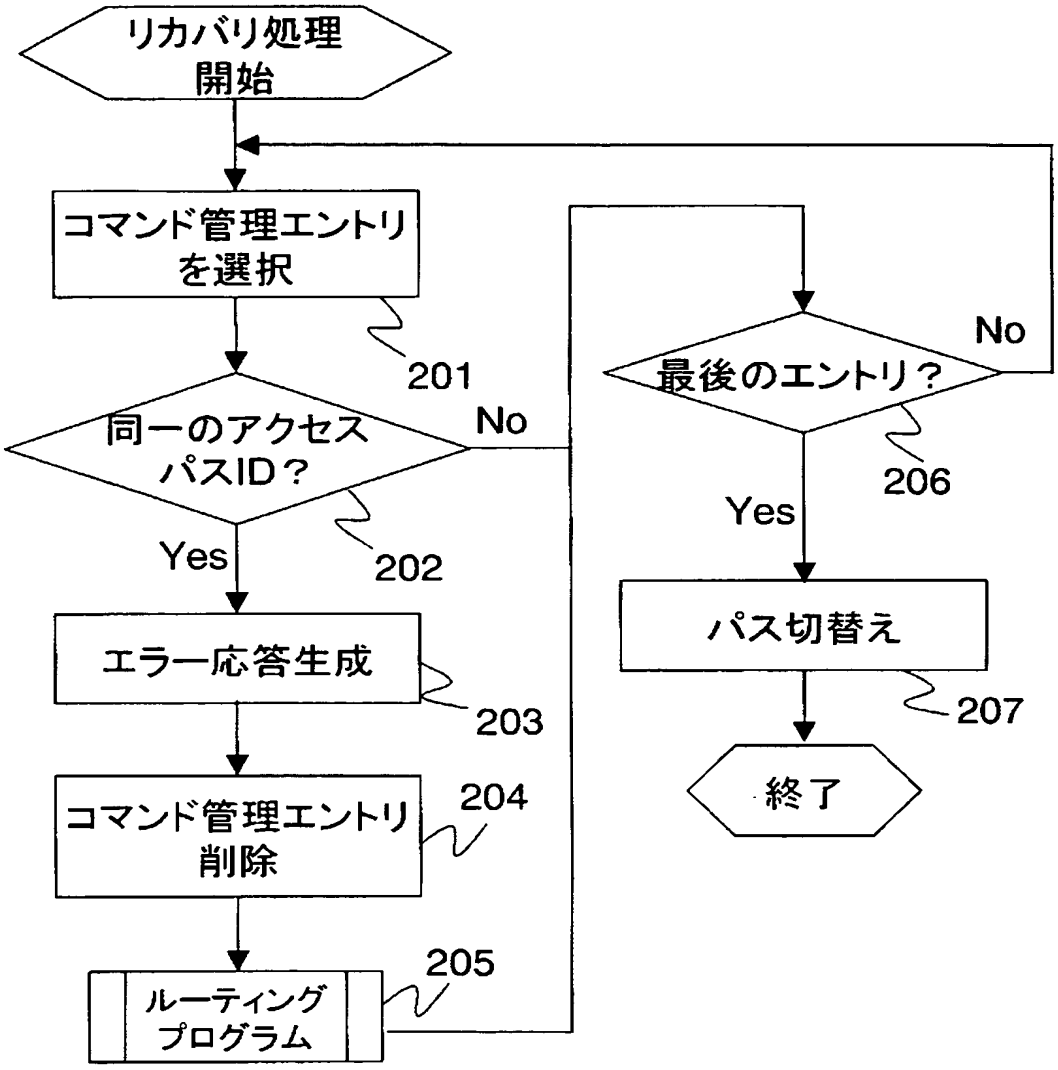
【図１３】

図 13


タイムアウト検出処理
開始
コマンド管理エントリ
を選択
191
有効なエントリ？
No
Yes
192
タイムアウトカウンタ
を1増やす
193
タイムアウト？
No
Yes
194
リカバリ
処理
195
最後のエントリ？
No
196
Yes
終了

【図１４】

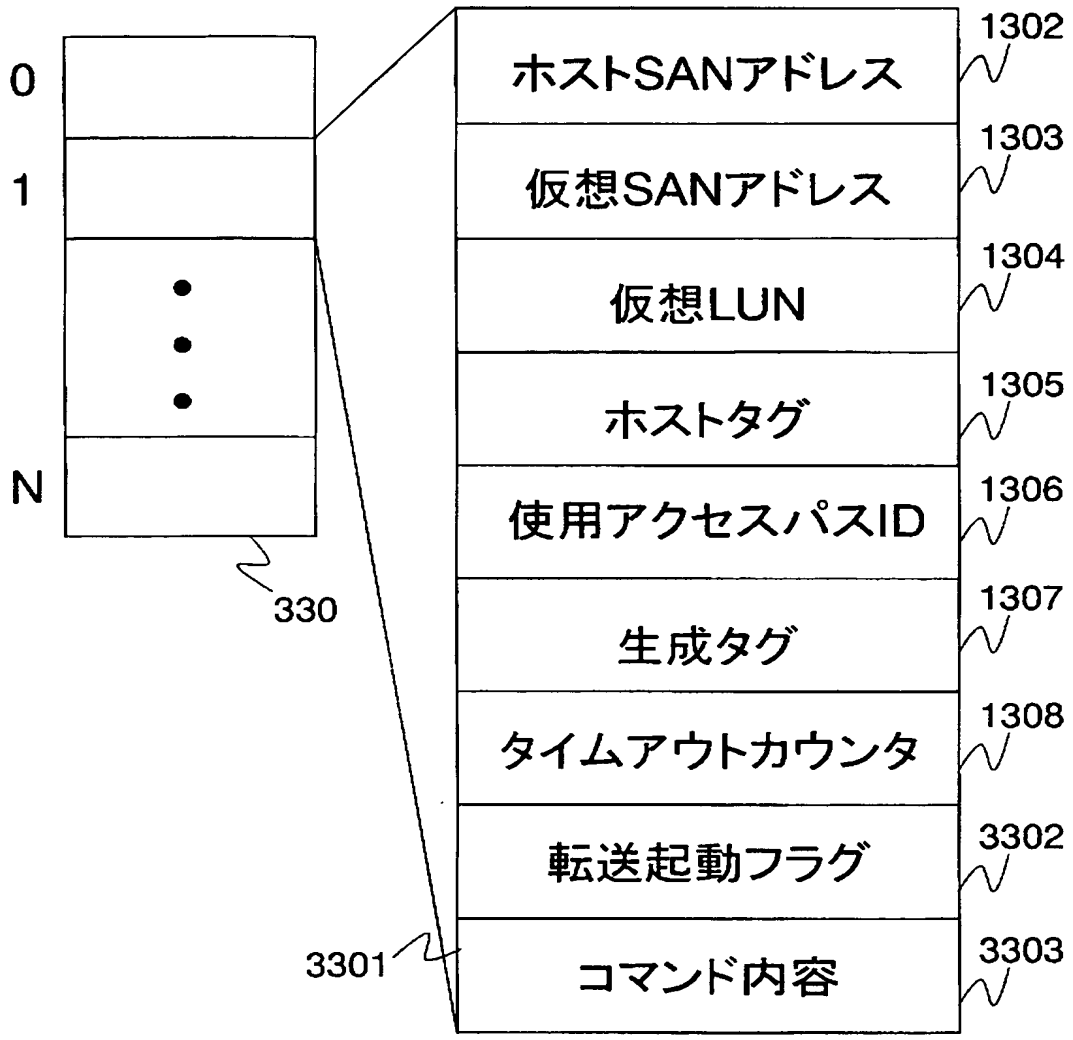
図 14


リカバリ処理
開始
コマンド管理エントリ
を選択
201
同一のアクセス
パスID？
No
202
Yes
エラー応答生成
203
コマンド管理エントリ
削除
204
ルーティング
プログラム
205
最後のエントリ？
No
206
Yes
パス切替え
207
終了

【図１５】

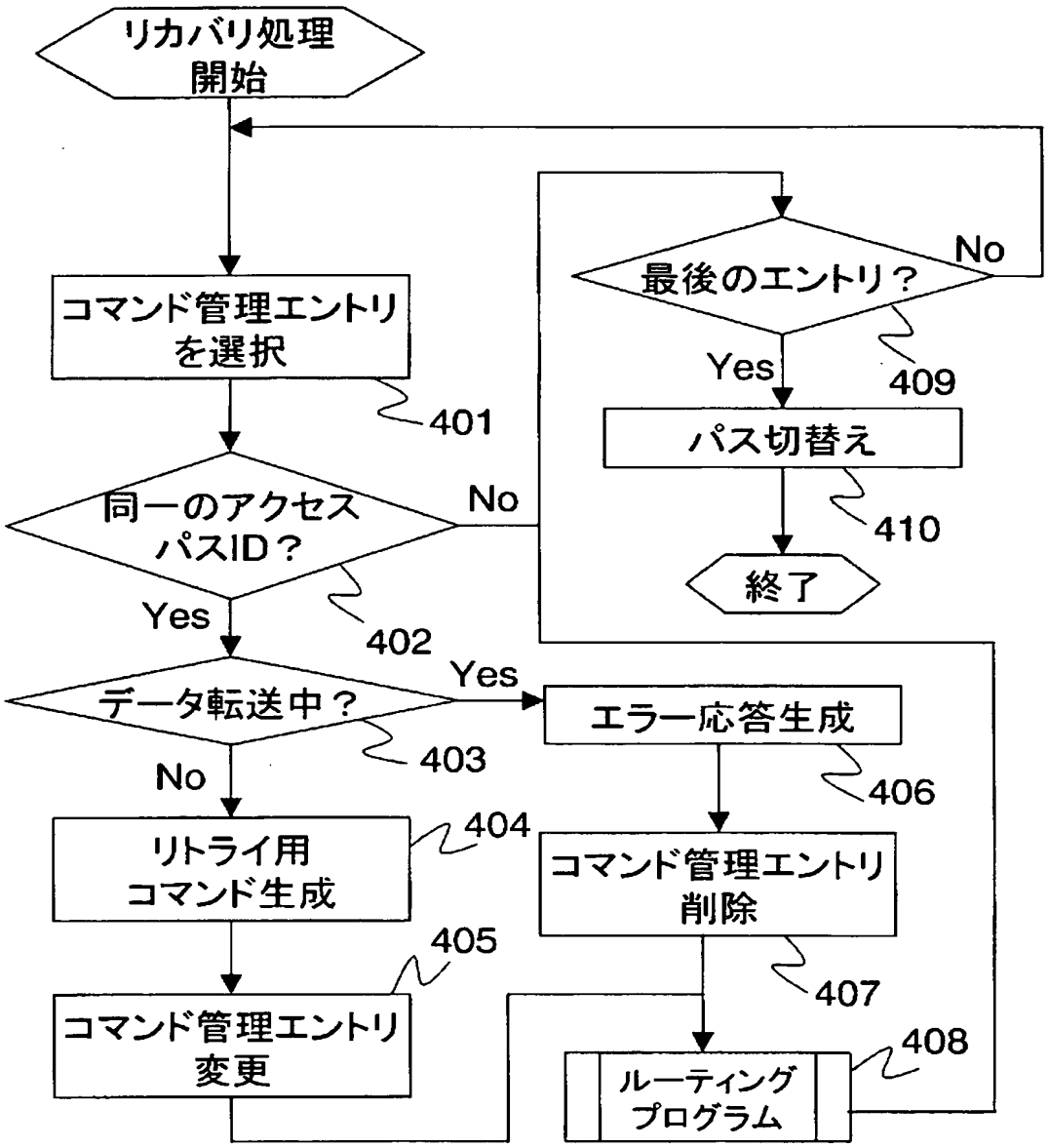
図 15

コマンド管理テーブル



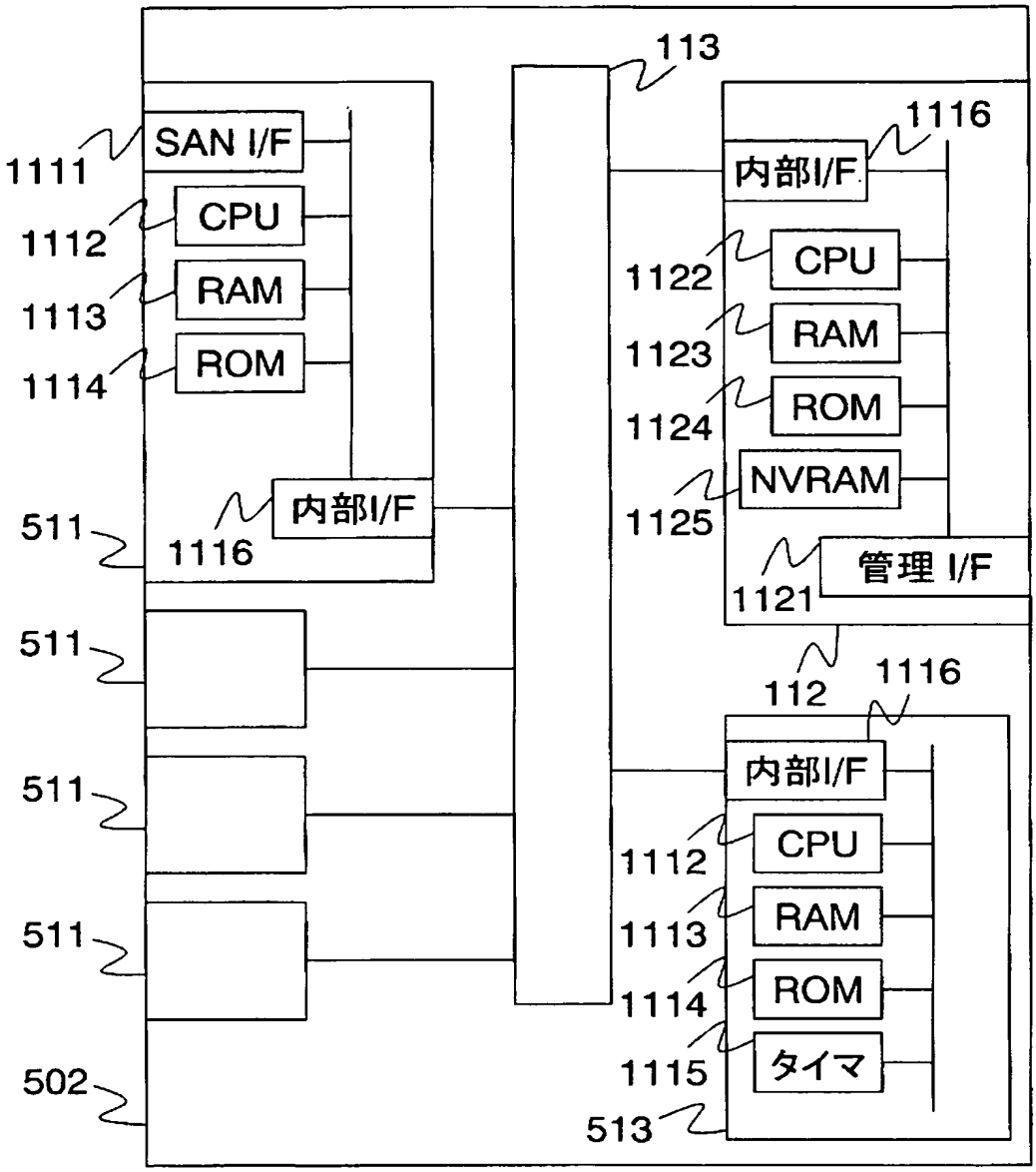
【図１６】

図 16


リカバリ処理
開始
コマンド管理エントリ
を選択
401
同一のアクセス
パスID？
No
Yes
402
データ転送中？
Yes
403
No
エラー応答生成
406
リトライ用
コマンド生成
404
コマンド管理エントリ
削除
407
コマンド管理エントリ
変更
405
ルーティング
プログラム
408
最後のエントリ？
No
Yes
409
パス切替え
410
終了

【図１７】

図 17


113
1116
SAN I/F
1111
内部I/F
CPU
1112
1122
CPU
RAM
1113
1123
RAM
ROM
1114
1124
ROM
NVRAM
1125
511
内部I/F
1116
管理 I/F
1121
511
112
1116
511
内部I/F
CPU
1112
511
1113
RAM
1114
ROM
1115
タイマ
502
513

【書類名】　要約書

【要約】

【課題】　従来の記憶装置の提供する経路障害の復旧手段はバッファないしキャッシュといったリソースが必要であり、従来のスイッチでは対応することが困難である。

【解決手段】　コンピュータと記憶装置の間のスイッチが、中継するコマンドを管理し、障害を検出するとエラー応答をコンピュータに送信すると同時に、以後のコマンドを代替経路に中継するよう設定を変更する。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００３－２０３４５４
受付番号	５０３０１２５７９７５
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　７月３１日

＜認定情報・付加情報＞

【提出日】　平成15年 7月30日

特願２００３－２０３４５４

出　願　人　履　歴　情　報

識別番号　　［０００００５１０８］

１．変更年月日　　１９９０年　８月３１日
［変更理由］　　新規登録
住　所　　東京都千代田区神田駿河台４丁目６番地
氏　名　　株式会社日立製作所